Adjusting for duration biases in sexual behaviour data

Jesse Knight[1,2] and Sharmistha Mishra[1,2]

[1]Institute of Medical Science, University of Toronto
[2]MAP Centre for Urban Health Solutions, Unity Health Toronto

July 17, 2023

Abstract

TODO

Funding.

Aknowledgements. We wish to thank: Saulius Simcikas, Jarle Tufto, and Michael Neely for help discussing and modelling biases. [1]

[1] See: `stats.stackexchange.com/questions/298828` and `math.stackexchange.com/questions/4732395`

1 Introduction

Epidemic modelling of sexually transmitted infections (STI) relies on quantification of sexual behaviour for model inputs (parameters) [1]. In models of STI transmission with risk heterogeneity — i.e., considering subgroups that experience differential risks — two important paramteters are: the duration of time within a "risk group", and the rate of sexual partnership formation, possibly stratified by partnership type [2–5]. For example, the average duration of time engaged in sex work can be used to define the modelled rate of "turnover" among sex workers [4]. Similarly, the numbers of main, casual, transactional, and/or paying sexual partners per year can be used to define the modelled rate of infection incidence [6].

Data to inform these parameters largely come from cross-sectional studies, and are often only available as crude aggregate estimates (vs individual-level data). Such estimates may be subject to distributional, sampling, censoring, and measurement biases. Our aim is therefore to develop bias adjuments for estimating:

1. duration in a risk group
2. rate of partnership change

from aggregate cross-sectional survey data, considering these factors. We explore these topics using aggregate estimates from a 2011 female sex worker survey in Eswatini [7].

2 Methods

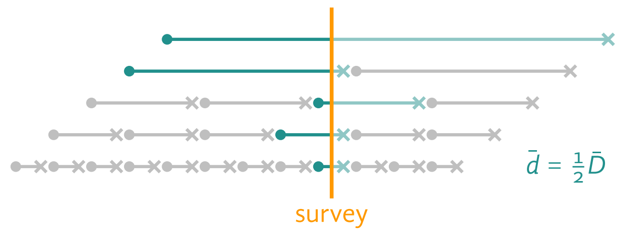
Data Source. Full details of the survey methodology are available in [8]. Briefly, 328 women aged 15+ who reported exchanging or selling sex for money, favors, or goods in the past 12 months were recruited via respondent-driven sampling (RDS) [9].

Approach. We conceptualize bias adjustments to the given data using Bayesian generative models — i.e., we define explicit distributions for the unbiased data and mechanisms of bias, and infer the parameters of these distributions based on the available data, using Gibbs sampling [10].[2] Implementation details are given in Appendix A.1.

2.1 Duration Selling Sex

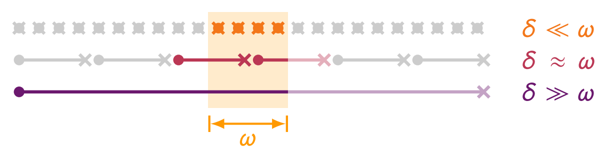
Crude Estimates. The survey [7] included questions about the current respondent's age and the age of first selling sex. The difference between these ages could be used to define a crude "duration selling sex". Using this approach, the crude median duration was $\tilde{d} = 4$ years. However, if durations are assumed to be exponentially distributed — a implicit assumption in compartmental models [11]

[2] We use JAGS: `mcmc-jags.sourceforge.io` via rjags: `cran.r-project.org/package=rjags`.

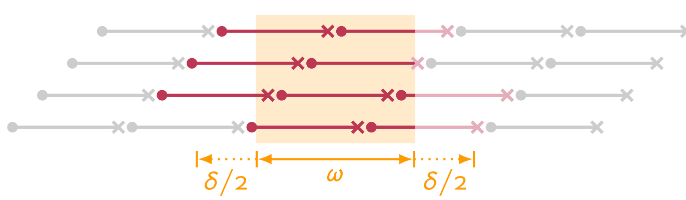


(a) Right censoring of reported durations selling sex in a steady state population

(b) Possible periods of selling sex for one individual who stopped 0, 1, or 2 times



(c) Differences in partnership duration vs recall period



(d) Fully and partially observed partnerships during a given recall period

Figure 1: Diagrams of fully observed, censored, and unobserved periods selling sex or within ongoing sexual partnerships

Guide: • : start, × : end, yellow: survey/recall period, full colour: fully observed, faded colour: right censored, grey: unobserved, $\bar{d}$: mean duration at survey and $\bar{D}$: overall, s: number of times stopped selling sex, g: relative gap length vs D, ω: recall period, δ: partnership duration, x: number of reported partnerships.

— then the crude mean could be estimated from the crude median as $\bar{d} = \tilde{d}/\log(2)$ due to skewness. Moving beyond crude estimates, we now develop the generative model considering the following potential biases.

Sampling. Sampling bias was considered via RDS-adjustment in [7], yielding mean and 95% CI estimates of the proportions of women p_z who had sold sex starting $\mathbb{d}_z \in \{0\text{–}2, 3\text{–}5, 6\text{–}10, 11+\}$ years ago (Table A.1). We define a model to identify distributions of reported durations d_i which are consistent with these data. We model each proportion p_z as a random variable with a beta approximation of binomial (BAB) distribution (see Appendix A.2) having parameters N_z and ρ_z. We model each N_z as a fixed value, which we fit to the 95% CI of p_z as described in § A.2. We then model each ρ_z as the proportion of reported durations d_i within the interval $\mathbb{d}_z$. Since these proportions are difficult to define analytically, we estimate $\hat{\rho}_z = \text{mean}(d_i \in \mathbb{d}_z)$ from $N = 328$ samples.

Censoring. These reported durations d_i are effectively right censored because they only capture engagement in sex work up until the survey, and and not additional sex work after the survey (Figure 1a) [12]. If we assume that the survey reaches women at a random time point during their total (eventual) duration selling sex D_i, we can model this censoring via a random fraction $f_i \sim \text{Unif}(0, 1)$, such that $d_i = f_i D_i$. The expected means are then related by $\bar{d}/\bar{D} = \bar{f} = \frac{1}{2}$. If we believe that the sampling adjustment above does not fully account for delays in self-identifying as a sex worker [13], we could instead use $f_i \sim \text{Unif}(1/4, 1)$, or similar.

Measurement. Finally, women may not sell sex continuously. Reported durations d_i may therefore include multiple periods of selling sex with gaps in between, whereas we aim to model D_i as the durations of individual periods selling sex. Respondents in [7] were not asked whether they ever stopped selling sex, but a later survey [14] indicated that $\phi = 45\%$ had stopped at least once. We model the number of times a woman stopped selling sex as a Poisson-distributed random variable s_i with mean $\bar{s}$. The expected value of ϕ given $\bar{s}$ is then $\text{P}(s > 0) = 1 - e^{-\bar{s}}$. Since $\phi = 45\%$ is an imperfect observation, we model ϕ as a random variable with a BAB distribution having parameters $N = 328$ and $\rho = 1 - e^{-\bar{s}}$, which allows inference on $\bar{s}$ given ϕ.

Next, we update the model for reported durations as $d_i = D_i(f_i + s_i(1 + g_i))$, where g_i is the relative duration of gaps between selling sex, with the following rationale. If $s_i = 0$, then $d_i = f_i D_i$ as before, reflecting the censored current period only. If $s_i > 0$, then d_i also includes s_i prior periods selling sex and the gaps between them (Figure 1b) — i.e., $s_i(D_i + g_i D_i) = D_i s_i(1 + g_i)$. The major assumption we make here is that all successive periods are of equal length, and likewise for gaps between them. We must also assume a distribution for g_i, for which we choose $g_i \sim \text{Exp}(1/\bar{g})$.

Summary. Figure 2a summarizes the proposed model graphically. The primary parameter of interest is the mean duration selling sex (for a given period) $\bar{D}$, but we must also infer the mean number of times women stop selling sex $\bar{s}$, and the mean relative duration of gaps $\bar{g}$. We assume uninformative priors for these 3 parameters.

2.2 Rates of Partnership Change

Data & Assumptions. The survey [7] also asked respondents to report their numbers of sexual partners (x) in a recall period (ω) of 30 days. Numbers were stratified by three types of partner: new paying clients, regular paying clients, and non-paying partners. We assume that only a small proportion of new clients go on to become regular clients; thus, we conceptualize "new" clients as effectively "one-off" clients.[3] Since no survey questions asked about partnership durations (δ), we further assume that these were: 1 day with new paying clients, 4 months with regular paying clients, and 3 years with non-paying partners. We now develop the generative model to estimate the expected rate of partnership change for each type, considering the following potential biases.

Sampling. As before, [7] estimates RDS-adjusted proportions of respondents p_z (mean, 95% CI) reporting different numbers/ranges of partners $\mathbb{x}_z$ in the past 30 days (Table A.1). Thus, we take the same approach as in § 2.1 to identify distributions of reported partner numbers x_i which are consistent with the data for each partnership type.

Interpretation. Numbers of reported partners (x) have generally been interpreted in two ways — x/ω as the *rate* of partnership change (Q) or x as the *number* of current partners (K):

$$Q \approx \frac{x}{\omega} \tag{13a}$$

or

$$K \approx x \tag{13b}$$

Both interpretations are reasonable under certain conditions: If partnership duration is short and the recall period is long ($\delta \ll \omega$), then reported partnerships mostly reflect *complete* partnerships, and thus $x/\omega \approx Q$. If partnership duration is long and the recall period is short ($\delta \gg \omega$), then reported partnerships mostly reflect *ongoing* partnerships, and thus $x \approx K$. However, if partnership duration and recall period are similar in length ($\delta \approx \omega$), then reported partnerships reflect a mixture of tail-ends, complete, and ongoing partnerships, and thus x/ω overestimates Q, but x also overestimates K. These three cases are illustrated in Figure 1c.

To adjust for this bias, we again assume that survey/recall period timing is effectively random. Then, if the *end* of the recall period would intersect an ongoing partnership, the intersection point would be, on average, half-way through the partnership. The same goes for the *start* of the recall period. Thus, the recall period is effectively extended by half the partnership duration $\delta/2$ on each end, and δ overall, as illustrated in Figure 1d. We can therefore define unbiased estimators of Q and K as:

$$Q = \frac{x}{\omega + \delta} \tag{14a}$$

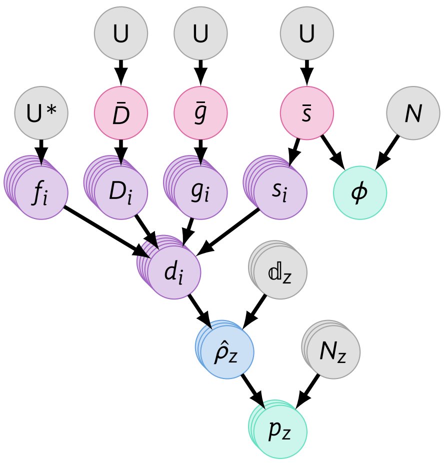
$$K = \frac{x\delta}{\omega + \delta} = Q\delta \tag{14b}$$

[3] The number of new clients per recall period could also be used to define a rate of partnership change [12], but we do not explore this approach here.

Returning to the generative model, we sample the true rate of partnership change from an assumed distribution $Q_i \sim \text{Gamma}(\alpha, \beta)$, with unknown parameters α, β. Then, we model the numbers of reported partners x_i given Q_i, ω, δ as $x_i \sim \text{Poi}(Q(\omega + \delta))$.

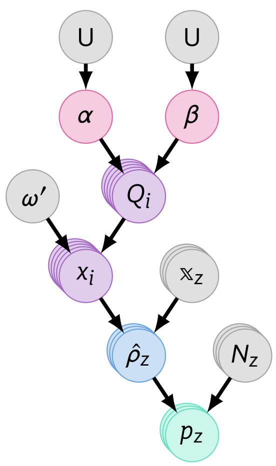
Summary. Figure 2b summarizes the proposed model graphically. The primary parameters of interest are α, β, which govern the distribution of rates of partnership change (for a given type) Q. We assume uninformative priors for these 2 parameters.

2.3 Experiment



$$p_z \sim \mathrm{BAB}\,(N_z, \hat{\rho}_z) \tag{1}$$
$$\hat{\rho}_z = \mathrm{mean}\,(d_i \in \mathbb{d}_z) \tag{2}$$
$$d_i = D_i\,(f_i + s_i\,(1 + g_i)) \tag{3}$$
$$D_i \sim \mathrm{Exp}\,(1/\bar{D}) \tag{4}$$
$$f_i \sim \mathrm{Unif}\,(0, 1) \tag{5}$$
$$s_i \sim \mathrm{Pois}\,(\bar{s}) \tag{6}$$
$$g_i \sim \mathrm{Exp}\,(1/\bar{g}) \tag{7}$$
$$\phi \sim \mathrm{BAB}\,(N, 1 - e^{-\bar{s}}) \tag{8}$$

(a) Duration selling sex



$$p_z \sim \mathrm{BAB}\,(N_z, \hat{\rho}_z) \tag{9}$$
$$\hat{\rho}_z = \mathrm{mean}\,(x_i \in \mathbb{x}_z) \tag{10}$$
$$x_i \sim \mathrm{Pois}\,(Q_i\,\omega') \tag{11}$$
$$Q_i \sim \mathrm{Gamma}\,(\alpha, \beta) \tag{12}$$

(b) Rates of partnership change

Figure 2: Graphical models

TODO

References

[1] G. P. Garnett. "An introduction to mathematical models in sexually transmitted disease epidemiology". In: *Sexually Transmitted Infections* 78.1 (Feb. 1, 2002), pp. 7–12. URL: https://doi.org/10.1136/sti.78.1.7.

[2] G. P. Garnett and R. M. Anderson. "Sexually transmitted diseases and sexual behavior: Insights from mathematical models". In: *Journal of Infectious Diseases* 174.S2 (Oct. 1, 1996), S150–S161. URL: https://doi.org/10.1093/infdis/174.Supplement˙2.S150.

[3] Hein Stigum, Per Magnus, and Leiv S. Bakketeig. "Effect of changing partnership formation rates on the spread of sexually transmitted diseases and human immunodeficiency virus". In: *American Journal of Epidemiology* 145.7 (Apr. 1, 1997), pp. 644–652. URL: https://10.0.4.69/oxfordjournals.aje.a009162.

[4] Charlotte Watts et al. "Remodelling core group theory: the role of sustaining populations in HIV transmission". In: *Sexually Transmitted Infections* 86.S3 (Dec. 1, 2010), pp. iii85–iii92. URL: https://doi.org/10.1136/sti.2010.044602.

[5] Jesse Knight et al. "Contribution of high risk groups' unmet needs may be underestimated in epidemic models without risk turnover: A mechanistic modelling analysis". In: *Infectious Disease Modelling* 5 (Jan. 1, 2020), pp. 549–562. URL: https://doi.org/10.1016/j.idm.2020.07.004.

[6] Marie-Claude Boily et al. "What really is a concentrated HIV epidemic and what does it mean for West and Central Africa? Insights from mathematical modeling". In: *Journal of Acquired Immune Deficiency Syndromes (1999)* 68 Suppl 2 (Mar. 1, 2015), S74–82. URL: https://doi.org/10.1097/QAI.0000000000000437.

[7] Stefan Baral et al. "Reconceptualizing the HIV epidemiology and prevention needs of female sex workers (FSW) in Swaziland". In: *PLOS ONE* 9.12 (Dec. 22, 2014), e115465. URL: http://doi.org/10.1371/journal.pone.0115465.

[8] Eileen A. Yam et al. "Association between condom use and use of other contraceptive methods among female sex workers in swaziland: A relationship-level analysis of condom and contraceptive use". In: *Sexually Transmitted Diseases* 40.5 (May 2013), pp. 406–412. URL: https://doi.org/10.1097/OLQ.0b013e318283c16d.

[9] Douglas D. Heckathorn. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations". In: *Social Problems* 44.2 (May 1, 1997), pp. 174–199. URL: https://doi.org/10.2307/3096941.

[10] Stuart Geman and Donald Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (Nov. 1984), pp. 721–741.

[11] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

[12] Erika Fazito et al. "Analysis of duration of risk behaviour for key populations: a literature review." In: *Sexually transmitted infections* 88.S2 (Dec. 1, 2012), pp. i24–i32. URL: https://doi.org/10.1136/sextrans-2012-050647.

[13] Eve Cheuk et al. "Transitions: Novel Study Methods to Understand Early HIV Risk Among Adolescent Girls and Young Women in Mombasa, Kenya, and Dnipro, Ukraine". In: *Frontiers in Reproductive Health* 2 (Sept. 10, 2020), p. 10. URL: https://doi.org/10.3389/frph.2020.00007.

[14] PEPFAR. *Characterizing the HIV Prevention and Treatment Needs among Key Populations, including Men who Have Sex with Men and Female Sex Workers in Swaziland: From Evidence to Action*. Mbabane, Swaziland, 2015.

A Supplement

A.1 Data & Code

Table A.1 gives the RDS-adjusted data from [7]. All analysis code is available online at: github.com/mishra-lab/duration-bias

Table A.1: RDS-adjusted proportions for variables

Variable	Stratum	Mean	(95% CI)
Years selling sex	0–2	38.3	(27.5, 49.1)
	3–5	32.1	(23.6, 40.7)
	6–10	20.2	(13.2, 27.1)
	11+	9.4	(04.4, 14.4)
New clients [a]	0–1	16.4	(09.8, 23.0)
	2	43.4	(33.3, 53.5)
	3	15.2	(09.6, 20.9)
	4	13.1	(07.0, 19.2)
	5	11.8	(06.0, 17.6)
Regular clients [a]	0–1	10.0	(01.9, 18.1)
	2	8.5	(03.2, 13.8)
	3	15.9	(09.8, 21.9)
	4	10.0	(04.5, 15.6)
	5	8.1	(03.8, 12.3)
	6	10.7	(05.8, 15.5)
	7+	36.9	(26.4, 47.3)
Non-paying partners [a]	0	12.5	(04.8, 20.1)
	1	50.8	(42.9, 58.7)
	2	23.6	(16.8, 30.3)
	3+	13.2	(07.2, 19.1)

[a] Number reported in the past 30 days. Data from [7].

A.2 Beta Approximation of the Binomial Distribution

The distributions of RDS-adjusted variables in [7] were reported as adjusted proportions (mean, 95% CI) for different variable value strata. For each proportion, we defined a beta approximation of the binomial (BAB) distribution:

$$\begin{aligned} \mathrm{P}(\rho) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\rho^{\alpha-1}(1-\rho)^{\beta-1} \qquad (A.1)\\ &\approx \binom{N}{n}\rho^{n}(1-\rho)^{N-n} \end{aligned}$$

with $\alpha = N\rho$ and $\beta = N(1-\rho)$. We fixed ρ as the adjusted point estimate, and estimated N by minimizing the sum of squared differences between the 95% quantiles of (A.1) given N and the reported 95% CI for the adjusted proportion.

A.3 Risk Group Duration

Fitting to RDS-Adjusted Proportions.

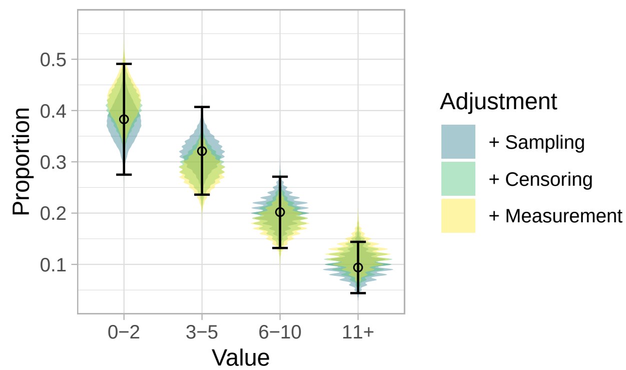


Figure A.1: TODO

TODO

Numeric Summary. Table A.2 summarizes the estimated exponential distribution means (95% CI) for years selling sex following each stage of adjustment outlined in § 2.1.

Table A.2: Estimated mean durations selling sex (years) following each stage of adjustment

Adjustment	Mean	(95% CI)
Median	4.00	—
Mean	5.77	—
+ Sampling	4.14	(3.44, 4.91)
+ Censoring	7.24	(5.65, 9.06)
+ Measurement	4.52	(3.40, 5.86)

A.4 Rate of Partnership Change

Fitting to RDS-Adjusted Proportions.

Numeric Summary.

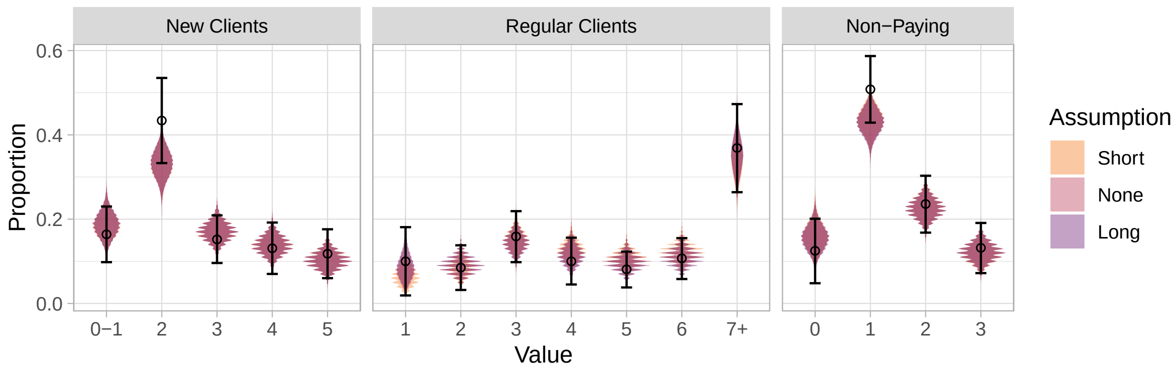


Figure A.2: TODO

TODO

Table A.3: Biased vs unbiased estimates of rates of partnership change and numbers of current partners for three partnership types reported by female sex workers

Partnership Type	Bias	Rate Q Mean	Rate Q (95% CI)	Number K Mean	Number K (95% CI)
New Clients	Biased	2.52	(2.26, 2.79)	2.52	(2.26, 2.79)
	Unbiased	2.44	(2.18, 2.70)	0.08	(0.07, 0.09)
Regular Clients	Biased	5.83	(5.12, 6.50)	5.83	(5.12, 6.50)
	Unbiased	1.17	(1.02, 1.30)	4.67	(4.10, 5.20)
Non-Paying	Biased	1.37	(1.19, 1.56)	1.37	(1.19, 1.56)
	Unbiased	0.04	(0.03, 0.04)	1.34	(1.16, 1.52)

Rares are per-month.