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 7, 2023

Abstract

TODO

Funding.

Aknowledgements. We wish to thank: Saulius Simcikas and Dr. Jarle Tufto for help identifying and discussing censoring biases.[1]

[1] stats.stackexchange.com/questions/298828

1 Introduction

Quantifying sexual behaviour is necessary to study the epidemiology of sexually transmitted infections (STI), including to inform inputs for STI transmission modelling [1]. Two important quantities for STI transmission modelling are: the duration of time within a “risk group” such as female sex workers (FSW), and the rate of partnership formation, possibly stratified by partnership type [2–5]. However, often only crude aggregate estimates of these quantities are available from previously published studies (vs individual-level data). Such estimates may be subject to distributional, sampling, censoring, and measurement biases.

Our aim is therefore to motivate and develop bias adjuments for estimating:

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

from cross-sectional survey data, considering these factors. We explore these topics using aggregate estimates from a 2011 FSW survey in Eswatini [6].

2 Methods

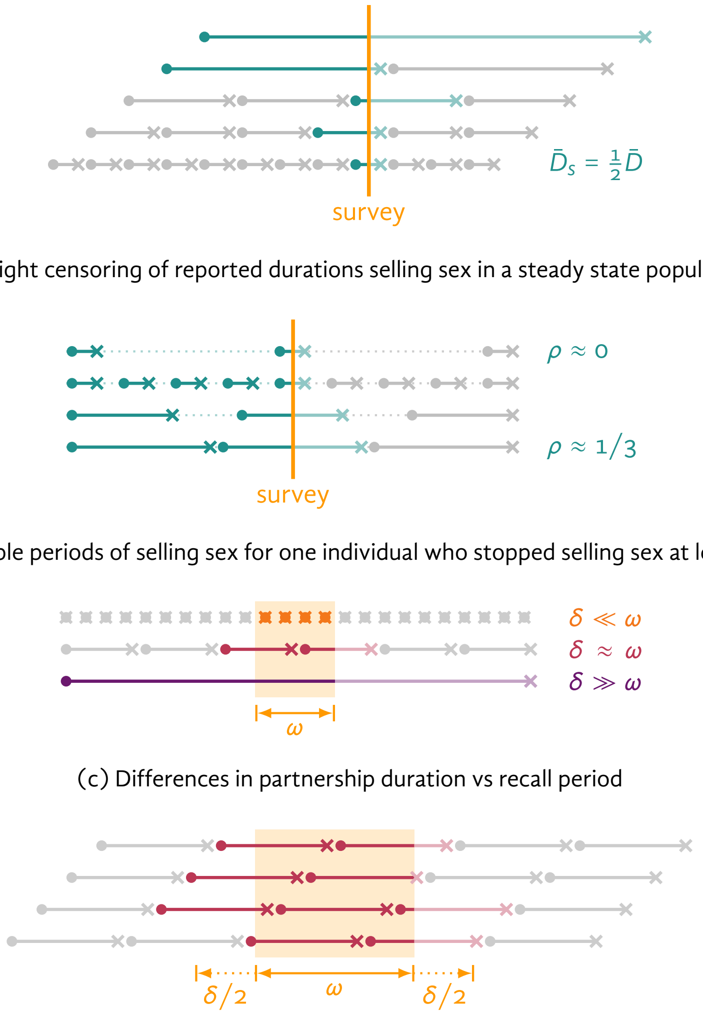
Data Source. Full details of the FSW survey methodology are available in [7]. Briefly, 328 women aged 15+ who reported exchanging or selling sex for money, favors, or goods in the past 12 months were recruited via respondent-driven sampling (RDS) [8].

2.1 Risk Group Duration

The FSW survey [6] 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 median unadjusted duration among FSW in Eswatini was 4 years. However, this estimate can be improved by considering the following potential biases.

Distribution. In compartmental transmission models, durations are implicitly assumed to be exponentially distributed [**?**]. This assumption was found to be reasonable here (see Figure A.1), but the median of an exponential distribution is less than the mean by a factor of log (2) due to skewness. Thus, the unadjusted mean duration could be estimated from the median as $\bar{D}_S = 4/\log(2) = 5.77$.

Sampling. Sampling error was considered via RDS-adjustment in [6], yielding estimates of the proportions of FSW who had sold sex starting 0–2, 3–5, 6–10, and 11+ years ago. The adjusted proportions indicate fewer years selling sex vs the unadjusted proportions, which would be consistent with challenges in reaching women in the first year(s) of sex work [9]. Fitting an exponential distribution to the cumulative adjusted proportions (methods in Appendix A.2; result in Figure A.1) yielded an estimated distribution mean $\bar{D}_S$ of 4.1 (95% CI: 3.4, 4.9) years.



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

(b) Possible periods of selling sex for one individual who stopped selling sex at least once

(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, green: selling sex, red: ongoing partnership, yellow: survey / recall period, colour: fully observed, faded colour: right censored, grey: unobserved, $\bar{D}_S$: and overall: $\bar{D}$mean duration at survey, ρ: proportion of total period in most recent period, ω: recall period, δ: partnership duration, x: number of reported partnerships.

Censoring. The reported durations are right censored because almost all respondents will continue selling sex after the survey (Figure 1a) [10]. If we assume that the survey reaches FSW at a random time point during their total (eventual) duration selling sex D, then the duration reported in the survey is effectively $D_S \sim \text{Unif}\,(0, D)$. Thus, the mean duration reported in the survey is $\bar{D}_S = \frac{1}{2}\bar{D}$, and we can define $f = \bar{D}/\bar{D}_S = 2$, to give a further adjusted estimate as: $\bar{D} = f\bar{D}_S$. In case the RDS-adjustment did not fully account for delayed self-identification as FSW, we could use $f \sim \text{Unif}\,(1.5, 2)$, or similar.

Measurement. Finally, FSW may not sell sex continuously. The 2011 survey did not ask whether respondents ever stopped selling sex, but a 2014 survey did, finding that $\phi = 45\%$ had stopped at least once [11]. Among the respondents who stopped, we have no further information about the proportion of the total period (i.e., since first started selling sex) reflected in the current period (i.e., since re-starting most recently). We denote this proportion ρ, and define the expected value for two extreme cases (Figure 1b top and bottom): respondents were almost *never* selling sex during the total period ($\rho = 0$), or respondents were almost *always* selling sex ($\rho = 1/3$). As shown in Figure 1b, the expected value for ρ given multiple stoppages is contained within these extremes. A strongly uninformative prior could then be $\rho \sim \text{Unif}\,(0, 1/3)$. Thus, we can define the final adjusted estimate for duration in sex work as: $\bar{D} = f[(1-\phi) + (\phi)\rho]\bar{D}_S$, with $\phi = 0.45$, $\rho \sim \text{Unif}\,(0, 1/3)$, and $f \sim \text{Unif}\,(1.5, 2)$.

2.2 Rates of Partnership Change

The FSW survey [6] also asked respondents to report their numbers of sexual partners in the past 30 days, stratified by three types of partner: new paying clients, regular paying clients, and non-paying partners. Our aim is to use the mean numbers of reported partners (x) for the 30-day recall period (ω), with the assumed partnership durations below (δ), to define expected rates of partnership change (Q). We also explore the expected number of current partners (K) below. We focus on means (not distributions), since rates of partnership change for each population are assumed to be homogeneous in compartmental transmission models. Given the assumptions below, we adjust for sampling and measurement bias to arrive at the final estimates.

Assumptions. For illustrative purposes, 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. Since no survey questions asked about partnership durations, we further assume that partnership durations were: 1 day with new paying clients, 4 months with regular paying clients, and 3 years with non-paying partners.

Sampling. Similar to years selling sex, RDS-adjusted proportions of respondents reporting different numbers of partners were given in [6]. Thus, following a similar approach as before (details in Appendix A.3), the mean numbers of reported partners of each type were estimated from these proportions as mean (95% CI): 2.52 (2.31, 2.75) new clients, 5.83 (5.15, 6.48) regular clients, and

1.38 (1.23, 1.52) non-paying partners (Figure A.2).

Rate or Number? Numbers of reported partners 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):

$$Q \approx \frac{x}{\omega} \tag{1a}$$

or

$$K \approx x \tag{1b}$$

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. Note that all estimates using (1) are biased via this same mechanism, just that some biases are larger than others.

Adjustment. To address this bias, we begin by assuming that survey timing is, as before, effectively random with respect to the durations of interest. Then, if either end of the recall period would capture an ongoing partnership, the intersection point would be, on average, at the partnership mid-point. 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. As such, we can define unbiased estimators of Q and K as:

$$Q = \frac{x}{\omega + \delta} \tag{2a}$$

$$K = \frac{x\delta}{\omega + \delta} = Q\delta \tag{2b}$$

To further reflect uncertainty in these estimates [TODO]. We then compared biased vs unbiased estimates of Q and K via (1) vs (2), for each partnership type.

Generalized Trends. To illustrate more general trends in the magnitude of bias, we further compared biased vs unbiased estimates of Q and K across a range of different partnership durations $\delta \in [0.1, 10]$ and recall periods $\omega \in [0.1, 10]$, with fixed true rate $Q = 1$ (arbitrary units).

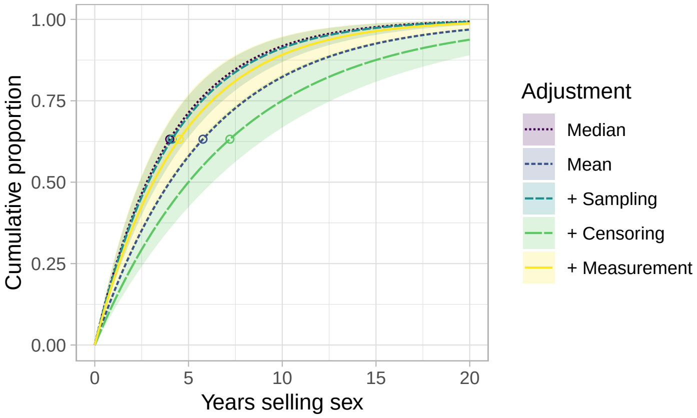


Figure 2: Estimated cumulative exponential distribution for years selling sex following each stage of adjustment

Guide: lines: distributon estimate, shaded ribbon: 95% CI, circles: distribution mean. Data from [6].

3 Results

3.1 Risk Group Duration

Figure 2 illustrates the estimated cumulative distributions for years selling sex following each stage of adjustment outlined in § 2.1; Table A.1 provides the corresponding distribution means (95% CI). Ironically, the final estimate of 4.52 was similar to the original median of 4, as each adjustment alternated betwen increasing and decreasing the estimated distribution mean. The censoring adjustment yielded the largest increase, while the measurement adjustment yielded the largest decrease.

3.2 Rates of Partnership Change

Figure 3 illustrates biased vs unbiased estimates of rates of partnership change (Q) and numbers of current partners (K), based on the RDS-adjusted numbers of reported partners (x) in 30 days (ω); Table A.2 provides the corresponding means (95% CI). The biased estimates of Q and K appear equal because Q is defined as per-month. We see that biases are strongest for Q with long partnerships (e.g., non-paying partners) and K with short partnership (e.g., new clients). However, biases are also substantial for both Q and K with "medium-length" partnerships (e.g., regular clients).

Figure 4 illustrates generalized trends in these biases, including 95% CI for unbiased estimates given different simulated survey sizes $N = 10, 100, 1000$.

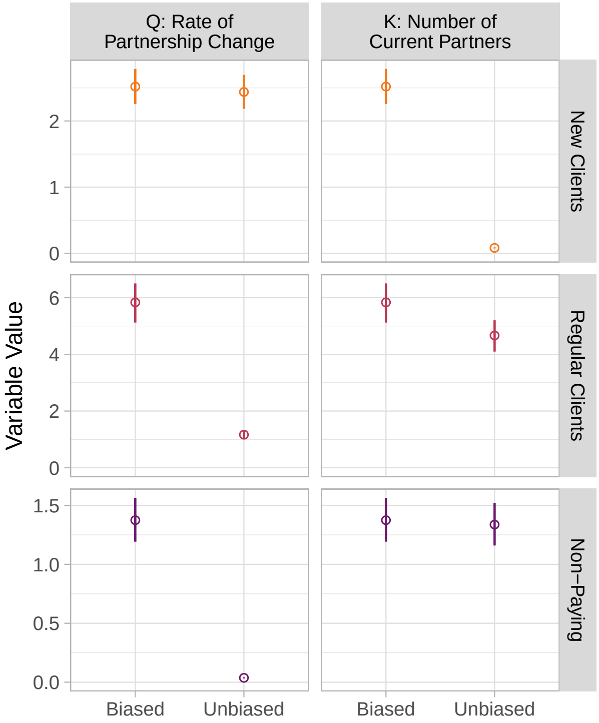


Figure 3: Biased vs unbiased estimates of rates of partnership change and numbers of current partners for three partnership types reported by female sex workers

Guide: circles: mean estimates, bars: 95% CI from 10,000 simulated surveys with $N = 328$. Rates are per-month. Data from [6].

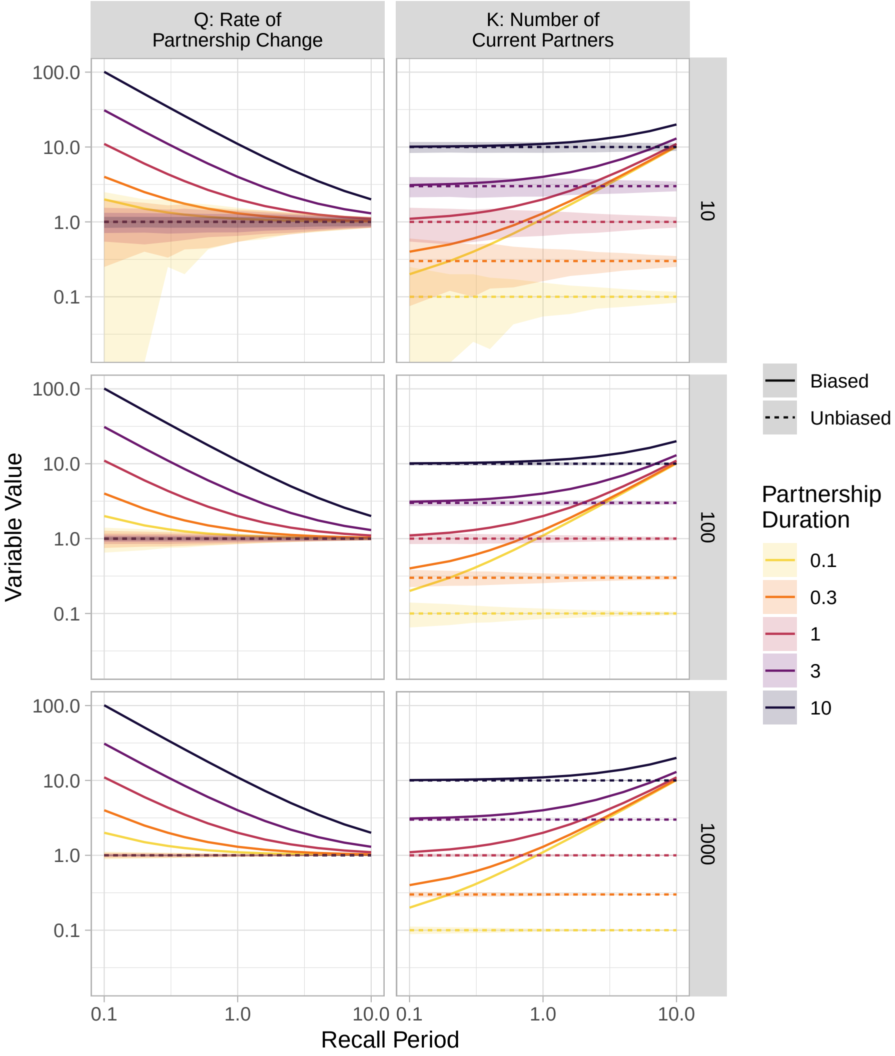


Figure 4: Biased vs unbiased estimates of rates of partnership change and numbers of current partners for different recall periods and partnership durations

Guide: shaded ribbon: 95% CI from 10,000 simulated surveys with N = 10, 100, 1000. Units are arbitrary.

4 Discussion

[TODO]

References

[1] K. A. Fenton et al. “Measuring sexual behaviour: Methodological challenges in survey research”. In: *Sexually Transmitted Infections* 77.2 (Apr. 1, 2001), pp. 84–92. URL: https://doi.org/10.1136/sti.77.2.84.

[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] Christopher J. Henry and James S. Koopman. “Strong influence of behavioral dynamics on the ability of testing and treating HIV to stop transmission”. In: *Scientific Reports* 5.1 (Aug. 22, 2015), p. 9467. URL: http://www.doi.org/10.1038/srep09467.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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

All analysis code is available online at: `github.com/mishra-lab/duration-bias`

A.1 Beta Approximation of the Binomial Distribution

The distributions of RDS-adjusted variables in [6] were reported as frequency tables with variable values and adjusted proportions (mean, 95% CI). For each proportion, we defined a beta approximation of the binomial (BAB) distribution:

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

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.2 Risk Group Duration

Fitted Exponential. We fit an exponential distribution (i.e., estimated λ) to the RDS-adjusted proportions for "years selling sex" in [6] by maximizing the overall likelihood (product of individual likelihoods) across BAB distributions, where each proportion was defined as:

$$\rho_i = F(t_i) - F(t_{i-1}), \quad F(t) = 1 - e^{-\lambda t} \tag{A.2}$$

Figure A.1 illustrates the resulting cumulative distribution vs the target cumulative proportions.

Numeric Summary. Table A.1 summarizes the estimated exponential distribution means (95% CI) for years selling sex following each stage of adjustment outlined in § 2.1.

Table A.1: 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)

Data from [6].

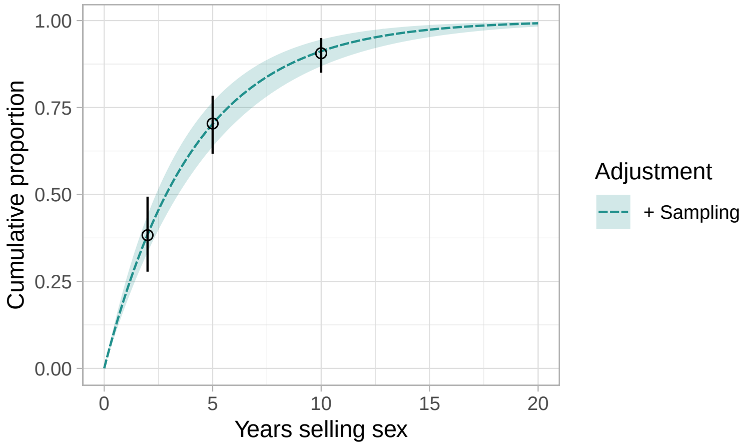


Figure A.1: Estimated cumulative exponential distribution for years selling sex, fitted to RDS-adjusted proportions

Guide: circles: mean adjusted cumulative proportions ρ_i selling sex for y_i years, bars: 95% CI for ρ_i, line: exponential distributon estimate, shaded ribbon: 95% CI. Data from [6]

A.3 Rate of Partnership Change

Mean Reported Partners. We repeated the following for each partnership type (new paying clients, regular paying clients, and non-paying partners). We estimated the mean number of reported partners $\bar{x}$ as a weighted average of the reported partners x_i with the RDS-adjusted proportions ρ_i:

$$\bar{x} = \sum_i \rho_i x_i \tag{A.3}$$

To capture uncertainty in the proportions ρ_i, we fitted a BAB distribution (i.e., estimated N_i) in each case, as described in § A.1. Then, we randomly sampled sets of ρ, normalized these sets by the sum (such that $\sum_i \rho_i = 1$), and re-calculated $\bar{x}$ for each set. Figure A.2 illustrates the resulting uncertainty distributions for $\bar{x}$, along with the corresponding x_i.

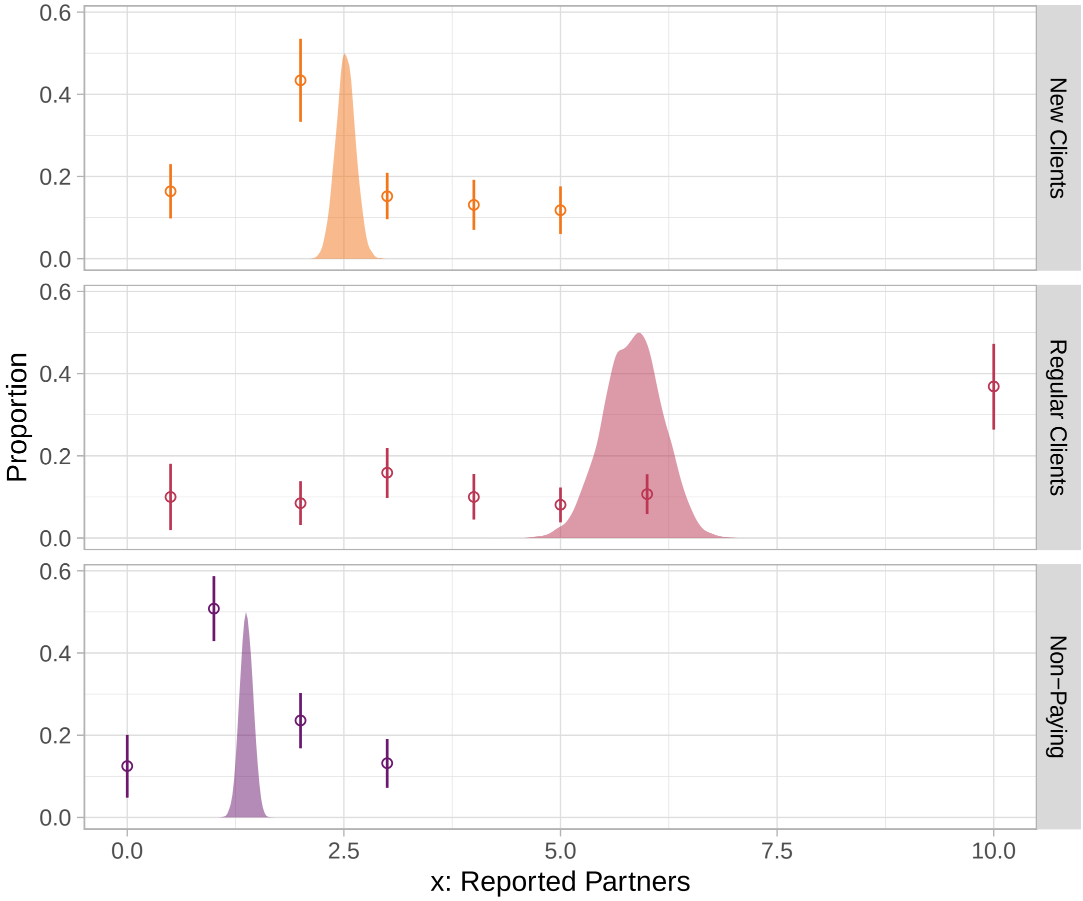


Figure A.2: Distributions of numbers of reported partners in the past 30 days

Guide: circles: mean adjusted proportions ρ_i reporting x_i partners, bars: 95% CI for ρ_i, shaded area: empiric distribution of $\bar{x}$. Data from [6].

Table A.2: Biased vs unbiased estimates of rates of partnership change and numbers of current partners for three partnership types reported by female sex workers

		Rate Q		Number K	
Partnership Type	Bias	Mean	(95% CI)	Mean	(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. Data from [6].