Observation Procedures and Markov Chain Models for Estimating the Prevalence and Incidence of a Behavior

James E. Pustejovsky
Northwestern University
pusto@u.northwestern.edu

Abstract

Data based on direct observation of behavior are used extensively in certain areas of educational and psychological research. A number of different procedures are used to record data during direct observation, including continuous recording, momentary time sampling (MTS), and partial interval recording (PIR). Among these, PIR has long been recognized as problematic because the mean of such data measures neither the prevalence nor the incidence of a behavior. However, little research has examined methods of analyzing PIR data other than simply summarizing it by the mean. I show that data collected using PIR can be represented using a discrete-time Markov chain derived from an alternating Poisson process model, which permits estimation of both prevalence and incidence via likelihood methods. Furthermore, I show that combining interval recording procedures with MTS considerably simplifies the Markov chain representation and leads to more precise estimates. Further work will study the operating characteristics of maximum likelihood estimators based on PIR data and on combined data, and will address questions of model fit for the underlying alternating Poisson process.

Background

- Direct observation of behavior is used extensively in certain areas of education and psychological research.
- Aspects of behavior that are often of interest include:
  - Prevalence: the proportion of time that a behavior occurs;
  - Incidence: the rate at which behavioral events occur.
- Procedures for recording observations vary in ease of implementation and level of detail.
  - Continuous recording methods are effort-intensive but produce rich data.
  - Less demanding methods are often required in applied settings.
  - Partial interval recording (PIR) is common but controversial because it does not directly measure prevalence or incidence (Kraemer, 1979; Rogosa & Ghandour, 1991).
- The relative merits of different methods remain open to debate, partly due to a lack of statistical work.

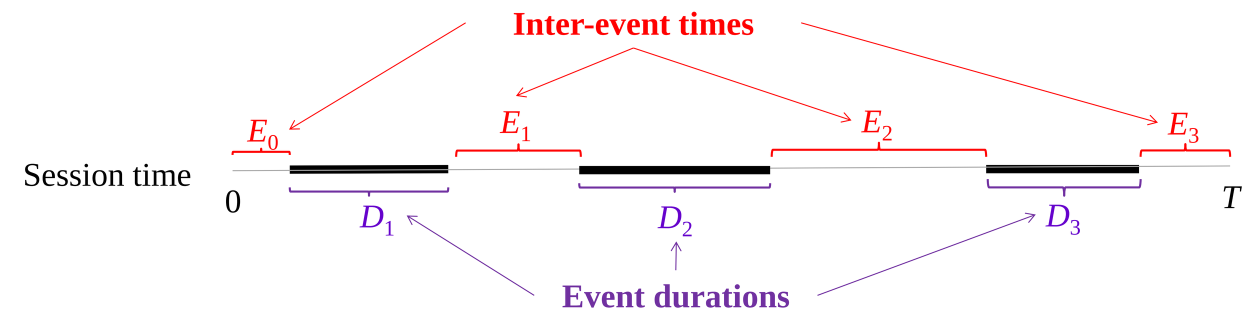
Goals

- Develop likelihood-based methods for estimating prevalence and incidence from PIR data.
- Develop new observation procedures that yield better estimates of these parameters.
- Provide guidance about appropriate circumstances for applying different observation procedures.

Literature cited

Brown, M., Solomon, H., & Stephens, M. A. (1977). Estimation of parameters of zero-one processes by interval sampling. *Operations Research*, *25*(3), 493–505.
Griffin, B., & Adams, R. (1983). A parametric model for estimating prevalence, incidence, and mean bout duration from point sampling. *American Journal of Primatology*, *4*(3), 261–271.
Kraemer, H. C. (1979). One-zero sampling in the study of primate behavior. Primates, 20(2), 237–244.
Rogosa, D., & Ghandour, G. (1991). Statistical Models for Behavioral Observations. *Journal of Educational Statistics*, *16*(3), 157–252.

Latent Alternating Poisson Process Model



Following Rogosa & Ghandour (1991), I use a model for the latent "behavior stream," from which the recorded data are derived. Over an observation of length T, assume that the behavior stream follows an alternating Poisson process where:

1. Event durations $D_1, D_2, D_3\ldots$ are independent and exponentially distributed with mean $\mathrm{E}(D_1) = \mu$;
2. Inter-event times $E_0, E_1, E_2, E_3\ldots$ are independent and exponentially distributed with mean $\mathrm{E}(E_1) = \lambda$;
3. Event durations and inter-event times are mutually independent;
4. The process is in equilibrium.

Under this model,

- $\phi = \mu / (\mu + \lambda)$ is the prevalence of the behavior and
- $\zeta = 1 / (\mu + \lambda)$ is the incidence of the behavior.

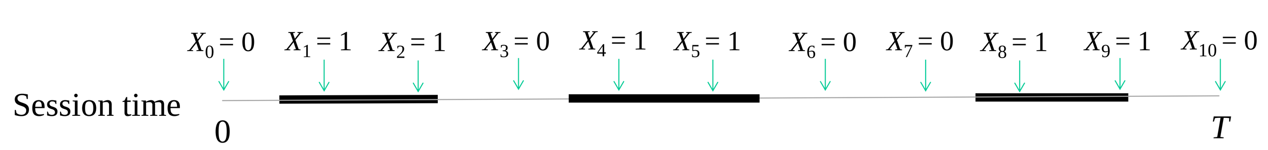
Let $Y(t)$ denote the state of the process at time t. Finally, define the transition probabilities:

$$p_0(t) = \Pr(Y(t) = 1 \mid Y(0) = 0) = \phi\left[1 - \exp\left(-\frac{\zeta t}{\phi(1-\phi)}\right)\right]$$

$$p_1(t) = \Pr(Y(t) = 1 \mid Y(0) = 1) = (1-\phi)\exp\left(-\frac{\zeta t}{\phi(1-\phi)}\right) + \phi$$

Momentary Time Sampling (MTS)

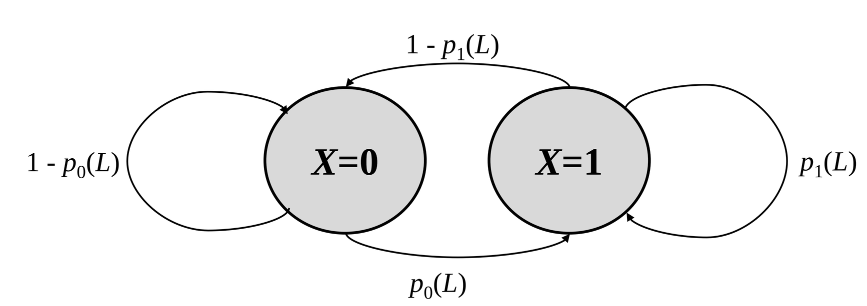
- Observer records the presence or absence of a behavior at each of $(K + 1)$ moments in time, equally spaced at intervals of length L.
- Recorded data are $X_k = Y(kL)$, $k = 0,\ldots,K$.



- Brown, Solomon, & Stephens (1977) demonstrate that, under an alternating Poisson process, MTS data follow a two-state discrete time Markov chain with transition probabilities

$$\Pr(X_k = 1 \mid X_{k-1} = a) = p_a(L) \quad \text{and} \quad \Pr(X_k = 0 \mid X_{k-1} = a) = 1 - p_a(L)$$

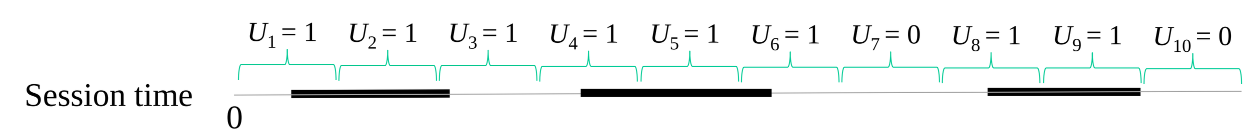
- Closed-form expressions for the maximum likelihood estimators for ϕ and γ are available (Brown, et al., 1977; Griffin & Adams, 1983).



Partial Interval Recording (PIR)

- Divide session into K intervals, each of length L.
- For each interval, observer records whether behavior occurred ***at any point*** during the interval.
- Recorded data are

$$U_k = I\left[0 < \int_{[0,L)} Y\left((k-1)L + t\right)dt\right], \quad k = 1,\ldots,K.$$



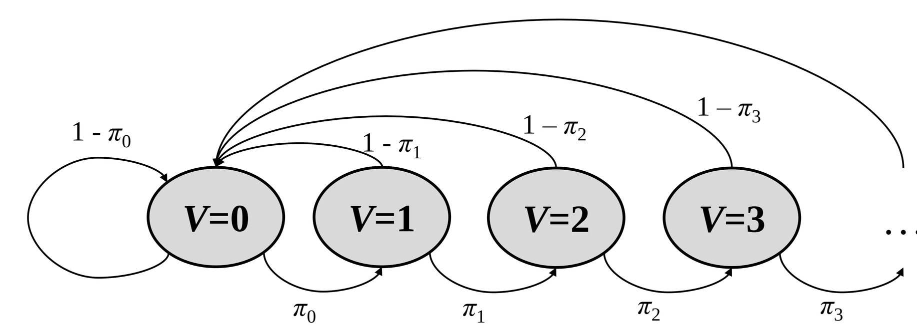
- Let $V_0 = 0$ and let V_k be the number of consecutive intervals where behavior is present: $V_k = k - \max\left\{0 \le j \le k : U_j = 0\right\}$.
- $V_1, V_2, \ldots, V_K$ form a discrete-time Markov chain on the space $\{0,1,2,3,\ldots\}$ with transition probabilities

$$\pi_j = \Pr\left(V_k = j+1 \mid V_{k-1} = j\right) = 1 - e^{-L/\lambda}\left[1 - f^{(j)}(0)\right]$$

where

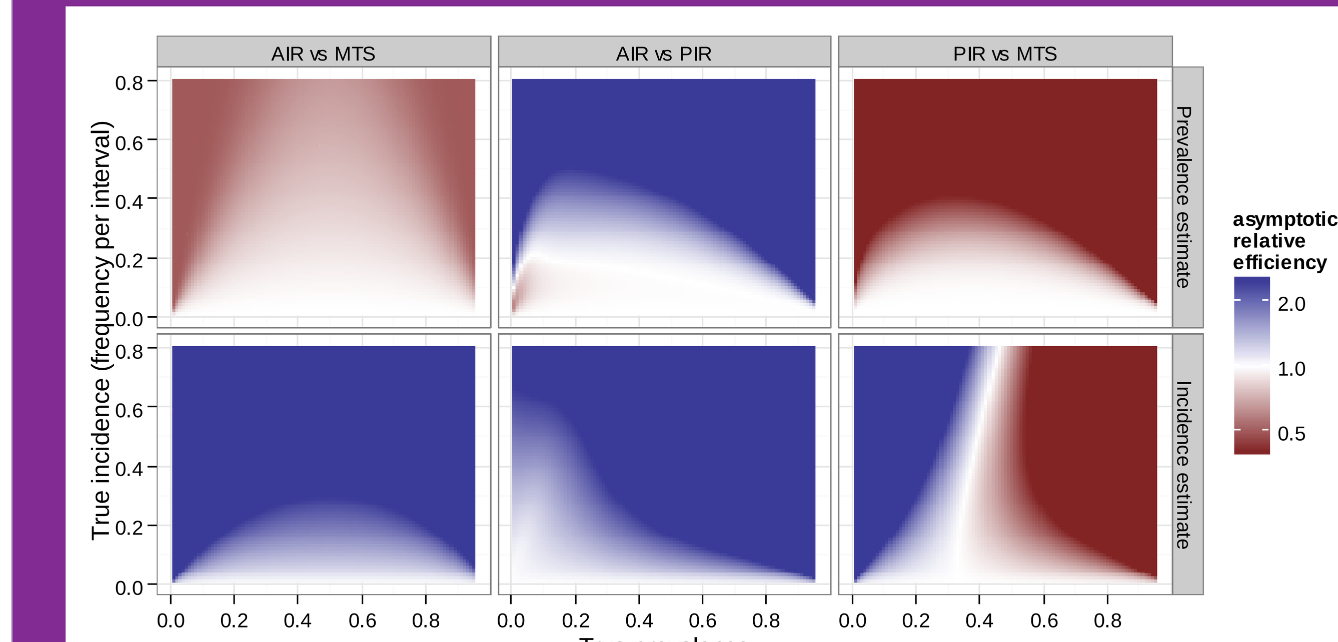
$$f(q) = \frac{\phi - (\phi - q)e^{-L\zeta/[\phi(1-\phi)]}}{1 - (1-q)e^{-L\zeta/(1-\phi)}}$$

and $f^{(j)}$ denotes the j-fold recursion of f.



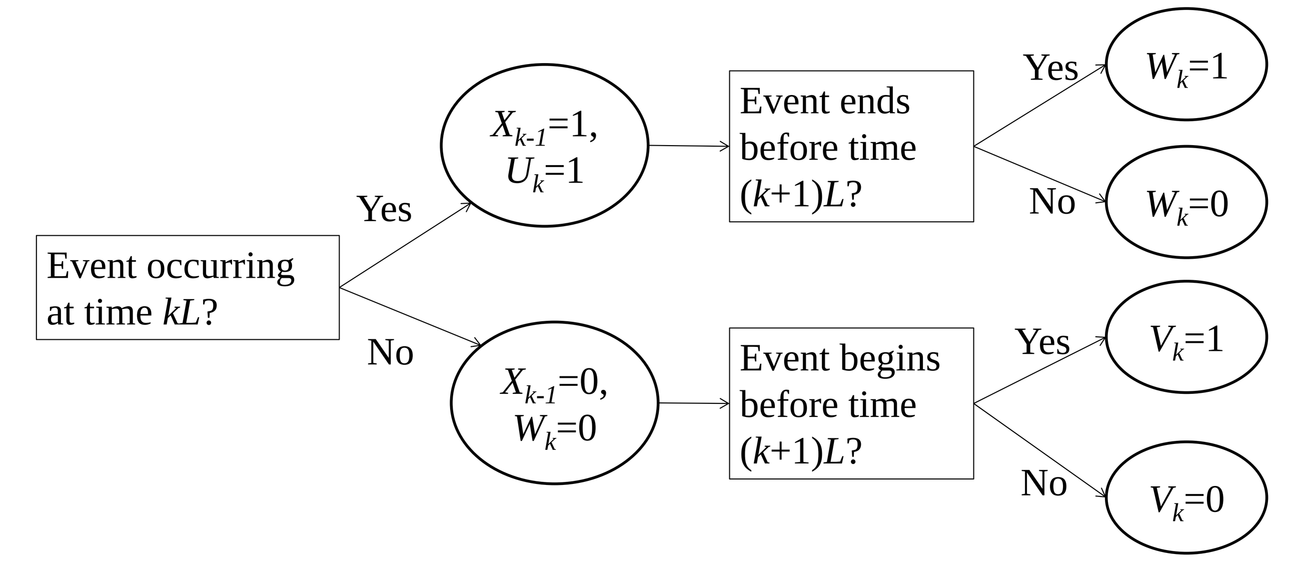
Whole interval recording (WIR)

- Divide session into K intervals, each of length L.
- For each interval, observer records whether behavior occurred ***for the duration*** of the interval.
- Recorded data are $W_k = I\left[L = \int_{[0,L)} Y\left((k-1)L + t\right)dt\right], \quad k = 1,\ldots,K.$
- Whole interval recording is equivalent to partial interval recording applied to the absence of a behavior rather than its presennce.
- The model for PIR data described above can therefore be used for either interval recording method, with appropriate interpretation of parameters.



Augmented Interval Recording (AIR)

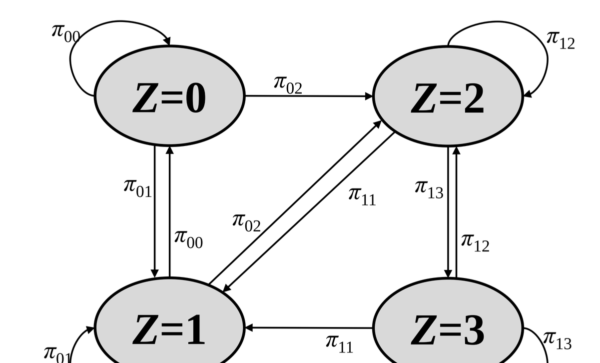
- MTS, PIR, and WIR can be combined as follows.
- Divide session into $K/2$ intervals, each of length $2L$.
- Use MTS at the beginning of each interval, to record X_{k-1}.
- If $X_{k-1} = 0$, use PIR for the remainder of the interval.
- If $X_{k-1} = 1$, use WIR for the remainder of the interval.
- Recorded data are $Z_k = X_k + U_k + W_k$.



- Conditional on X_0, $(Z_1,\ldots,Z_{K/2})$ form a discrete-time Markov chain on the space $\{0,1,2,3\}$, with transition probabilities $\pi_{ab} = \Pr(Z_k = b \mid X_{k-1} = a)$ for

$$\pi_{00} = e^{-2L/\lambda} \quad \pi_{01} = 1 - e^{-2L/\lambda} - p_0(2L) \quad \pi_{02} = p_0(2L) \quad \pi_{03} = 0$$

$$\pi_{10} = 0 \quad \pi_{11} = \frac{\lambda}{\mu}p_0(2L) \quad \pi_{12} = 1 - e^{-2L/\lambda} - \frac{\lambda}{\mu}p_0(2L) \quad \pi_{13} = e^{-2L/\lambda}$$



Asymptotic Relative Efficiency

- For a recording procedure following a DTMC on an $(R+1)$-dimensional space, the expected information from K observations has $(i,j)^{\text{th}}$ entry

$$\left[I^{(K)}\right]_{ij} = K\sum_{r=0}^{R}\sum_{q=0}^{R}\frac{\tilde{\pi}_q}{\pi_{qr}}\frac{\partial \pi_{qr}}{\partial \theta_i}\frac{\partial \pi_{qr}}{\partial \theta_j}$$

where $\boldsymbol{\theta} = (\phi,\zeta)^{\mathrm{T}}$ and $\tilde{\pi}_q$ is the limiting occupancy probability in state q.

- For parameter θ_i, the asymptotic relative efficiency of the procedures is calculated as
  - AIR versus MTS: $\left[I_{MTS}^{(2)}\right]_{ii}^{-1} / \left[I_{AIR}^{(1)}\right]_{ii}^{-1}$
  - AIR versus PIR: $\left[I_{PIR}^{(2)}\right]_{ii}^{-1} / \left[I_{AIR}^{(1)}\right]_{ii}^{-1}$
  - PIR versus MTS: $\left[I_{MTS}^{(1)}\right]_{ii}^{-1} / \left[I_{PIR}^{(1)}\right]_{ii}^{-1}$
- The relative efficiency of the procedures is plotted in the figure to the left for various values of prevalence and incidence.
- Over most of the parameter space, AIR dominates PIR for both prevalence and incidence.