Fisher Information, stochastic processes and generating functions

Ali Eshragh

(Joint work with Nigel Bean, Joshua Ross and Bruno Salvy)

School of Mathematical and Physical Sciences & CARMA
The University of Newcastle, Australia

CARMA Workshop in Honour of Brailey Sims
August, 2015 - Newcastle

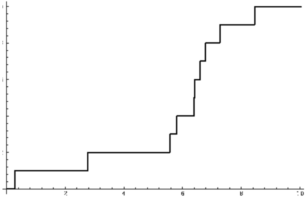
Motivation

- **Epidemiology**

Motivation

- **Epidemiology**

- A **Growing** Population



Definition and Notation

- Let X_t denote the **population size** at time t.

Definition and Notation

- Let X_t denote the **population size** at time t.
- $\{X_t : t \in \mathrm{R}_0^+\}$ is a **stochastic process** .

Definition and Notation

- Let X_t denote the **population size** at time t.
- $\{X_t : t \in \mathrm{R}_0^+\}$ is a **stochastic process** .
- Suppose $\{X_t : t \in \mathrm{R}_0^+\}$ is a **simple birth process (SBP)** with the **birth rate** λ. Moreover, $X_0 \overset{a.s.}{=} x_0$.

Definition and Notation

- Let X_t denote the **population size** at time t.
- $\{X_t : t \in \mathrm{R}_0^+\}$ is a **stochastic process** .
- Suppose $\{X_t : t \in \mathrm{R}_0^+\}$ is a **simple birth process (SBP)** with the **birth rate** λ. Moreover, $X_0 \overset{a.s.}{=} x_0$.
- It is **Markovian**, that is

$$\Pr(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n, \ldots, , X_{t_1} = x_1) \;=\; \Pr(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n) \,,$$

for all possible values of n and $t_1, \ldots, t_{n+1}$.

Definition and Notation

- Let X_t denote the **population size** at time t.
- $\{X_t : t \in \mathrm{R}_0^+\}$ is a **stochastic process** .
- Suppose $\{X_t : t \in \mathrm{R}_0^+\}$ is a **simple birth process (SBP)** with the **birth rate** λ. Moreover, $X_0 \stackrel{a.s.}{=} x_0$.
- It is **Markovian**, that is

$$\Pr(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n, \ldots, , X_{t_1} = x_1) = \Pr(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n),$$

  for all possible values of n and $t_1, \ldots, t_{n+1}$.
- The **transition probability** is equal to

$$\Pr(X_{s+t} = j | X_s = i) = \binom{j-1}{i-1} e^{-\lambda t i}(1 - e^{-\lambda t})^{j-i} .$$

Likelihood Function

- **Estimating** the unknown parameter λ through **maximum likelihood** method.

Likelihood Function

- **Estimating** the unknown parameter λ through **maximum likelihood** method.
- Take the **observations** $X_{t_1}, \ldots, X_{t_n}$ at observation times $0 < t_1 \leq \ldots \leq t_n \leq \tau$, respectively.

Likelihood Function

- **Estimating** the unknown parameter λ through **maximum likelihood** method.
- Take the **observations** $X_{t_1}, \ldots, X_{t_n}$ at observation times $0 < t_1 \leq \ldots \leq t_n \leq \tau$, respectively.
- Construct the **likelihood function**

$$\mathcal{L}(x_1, \ldots, x_n; \lambda) = \Pr(X_{t_1} = x_1, \ldots, X_{t_n} = x_n | \lambda)$$

Likelihood Function

- **Estimating** the unknown parameter λ through **maximum likelihood** method.
- Take the **observations** $X_{t_1}, \ldots, X_{t_n}$ at observation times $0 < t_1 \leq \ldots \leq t_n \leq \tau$, respectively.
- Construct the **likelihood function**

$$\begin{aligned}\mathcal{L}(x_1, \ldots, x_n; \lambda) &= \Pr(X_{t_1} = x_1, \ldots, X_{t_n} = x_n | \lambda) \\ &= \prod_{i=1}^{n} \binom{x_i - 1}{x_{i-1} - 1} e^{-\lambda(t_i - t_{i-1})x_{i-1}} \left(1 - e^{-\lambda(t_i - t_{i-1})}\right)^{x_i - x_{i-1}}.\end{aligned}$$

Observation Times

- **When** should we take the observations $X_{t_1}, \ldots, X_{t_n}$?

Observation Times

- **When** should we take the observations $X_{t_1}, \ldots, X_{t_n}$?
- Presumably, a good choice is finding observation times $t_1, \ldots, t_n$ such that the **expected volume of information** obtained from these observations to estimate the unknown parameter λ is **maximized**.

Observation Times

- **When** should we take the observations $X_{t_1}, \ldots, X_{t_n}$?
- Presumably, a good choice is finding observation times $t_1, \ldots, t_n$ such that the **expected volume of information** obtained from these observations to estimate the unknown parameter λ is **maximized**.
- A good tool to measure the expected volume of information gained from a set of observations is the **Fisher Information**.

Observation Times

- **When** should we take the observations $X_{t_1}, \ldots, X_{t_n}$?
- Presumably, a good choice is finding observation times $t_1, \ldots, t_n$ such that the **expected volume of information** obtained from these observations to estimate the unknown parameter λ is **maximized**.
- A good tool to measure the expected volume of information gained from a set of observations is the **Fisher Information**.
- It can be shown that

$$\mathcal{FI}_{(X_{t_1},\ldots,X_{t_n})}(\lambda) = E_{\mathcal{L}}\left[\left(\frac{d}{d\lambda}\ln(\mathcal{L}(X_{t_1},\ldots,X_{t_n};\lambda))\right)^2\right].$$

Observation Times

- **When** should we take the observations $X_{t_1}, \dots, X_{t_n}$?
- Presumably, a good choice is finding observation times $t_1, \dots, t_n$ such that the **expected volume of information** obtained from these observations to estimate the unknown parameter λ is **maximized**.
- A good tool to measure the expected volume of information gained from a set of observations is the **Fisher Information**.
- It can be shown that

$$\mathcal{FI}_{(X_{t_1},\dots,X_{t_n})}(\lambda) = E_{\mathcal{L}}\left[\left(\frac{d}{d\lambda}\ln(\mathcal{L}(X_{t_1},\dots,X_{t_n};\lambda))\right)^2\right].$$

- Hence, $(t_1^*, \dots, t_n^*) \in \text{argmax}\{\mathcal{FI}_{(X_{t_1},\dots,X_{t_n})}(\lambda)\}$.

Fisher Information and Optimal Observation Times

Proposition (Becker and Kersting, 1983)

The **Fisher information** *for a SBP with the parameter λ, the initial value of x_0 and the observation times of $(t_1, \ldots, t_n)$ is as follows:*

$$\mathcal{FI}_{(x_{t_1},\cdots,x_{t_n})}(\lambda) \quad = \quad x_0 \sum_{i=1}^{n} \frac{(t_i - t_{i-1})^2}{e^{-\lambda t_{i-1}} - e^{-\lambda t_i}} \, .$$

Fisher Information and Optimal Observation Times

Proposition (Becker and Kersting, 1983)

The **Fisher information** *for a SBP with the parameter* λ, *the initial value of* x_0 *and the observation times of* $(t_1, \ldots, t_n)$ *is as follows:*

$$\mathcal{FI}_{(X_{t_1}, \cdots, X_{t_n})}(\lambda) = x_0 \sum_{i=1}^{n} \frac{(t_i - t_{i-1})^2}{e^{-\lambda t_{i-1}} - e^{-\lambda t_i}} .$$

Optimal Observation Times (Becker and Kersting, 1983)

$$\mathbf{t_i^* \approx \frac{3}{\lambda} \log\left(1 + \frac{i}{n}(e^{\frac{\lambda\tau}{3}} - 1)\right) \quad for\ i = 1, \ldots, n}$$

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.
- At each observation time, each individual can be counted **independently** with probability **p**.

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.
- At each observation time, each individual can be counted **independently** with probability **p**.
- $\mathbf{Y_t}$ is the number of individuals observed at at time t.

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.
- At each observation time, each individual can be counted **independently** with probability $\mathbf{p}$.
- $\mathbf{Y_t}$ is the number of individuals observed at at time t.
- $(Y_t|X_t = x) \sim \texttt{Binomial}(\mathbf{x}, \mathbf{p})$.

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.
- At each observation time, each individual can be counted **independently** with probability $\mathbf{p}$.
- $\mathbf{Y_t}$ is the number of individuals observed at at time t.
- $(Y_t|X_t = x) \sim \mathtt{Binomial}(\mathbf{x}, \mathbf{p})$.
- We call the stochastic process $\{Y_t : t \in \mathrm{R}_0^+\}$ the **partially-observable simple birth process (POSBP)** with parameters (λ, p).

Definition and Notation

- Suppose that at each observation time, we can count the population, **partially**.
- At each observation time, each individual can be counted **independently** with probability **p**.
- $\mathbf{Y_t}$ is the number of individuals observed at at time t.
- $(Y_t|X_t = x) \sim \texttt{Binomial}(\mathbf{x}, \mathbf{p})$.
- We call the stochastic process $\{Y_t : t \in \mathrm{R}_0^+\}$ the **partially-observable simple birth process (POSBP)** with parameters (λ, p).
- $\text{POSBP}(\lambda, 1) \equiv \text{SBP}(\lambda)$.

Markovian or non-Markovian?

Theorem (Bean, Elliott, Eshragh and Ross; 2015)

The POSBP $\{Y_t : t \in \mathrm{R}_0^+\}$ *with parameters* (λ, p) *is* **not Markovian**.

Markovian or non-Markovian?

Theorem (Bean, Elliott, Eshragh and Ross; 2015)

The POSBP $\{Y_t : t \in \mathrm{R}_0^+\}$ *with parameters* (λ, p) *is* **not Markovian**.

- However,

$$\Pr(Y_{t_1} = y_{t_1}, \ldots, Y_{t_n} = y_{t_n} | X_{t_1} = x_{t_1}, \ldots, X_{t_n} = x_{t_n})$$

$$= \prod_{i=1}^{n} \Pr(Y_{t_i} = y_{t_i} | X_{t_i} = x_{t_i}) .$$

Likelihood Function

- The likelihood function:

$$\mathcal{L}(y_{t_1}, \ldots, y_{t_n};\ \lambda, p) \ = \ \Pr(Y_{t_1} = y_{t_1}, \ldots, Y_{t_n} = y_{t_n})$$

Likelihood Function

- The likelihood function:

$$
\begin{aligned}
\mathcal{L}(y_{t_1}, \ldots, y_{t_n}; \lambda, p) &= \Pr(Y_{t_1} = y_{t_1}, \ldots, Y_{t_n} = y_{t_n}) \\
&= \sum_{x_{t_1}, \ldots, x_{t_n}} \prod_{i=1}^{n} \binom{x_{t_i}}{y_{t_i}} p^{y_i} q^{x_{t_i} - y_{t_i}} \binom{x_{t_i} - 1}{x_{t_{i-1}} - 1} \upsilon_{i-1,i}^{x_{t_{i-1}}} (1 - \upsilon_{i-1,i})^{x_{t_i} - x_{t_{i-1}}},
\end{aligned}
$$

where $q := 1 - p$ and $\upsilon_{i-1,i} := e^{-\lambda(t_i - t_{i-1})}$.

Truncated Summation

- Fisher Information:

$$\mathcal{FI}_{(Y_{t_1},\ldots,Y_{t_n})}(\lambda) \quad = \quad \sum_{y_{t_n}=0}^{\infty} \cdots \sum_{y_{t_1}=0}^{\infty} \frac{\left(\frac{d\mathcal{L}(y_{t_1},\ldots,y_{t_n};\,\lambda,p)}{d\lambda}\right)^2}{\mathcal{L}(y_{t_1},\ldots,y_{t_n};\;\lambda,p)}.$$

Truncated Summation

- Fisher Information:

$$\mathcal{FI}_{(Y_{t_1},\ldots,Y_{t_n})}(\lambda) \;=\; \sum_{y_{t_n}=0}^{\infty} \cdots \sum_{y_{t_1}=0}^{\infty} \frac{\left(\frac{d\mathcal{L}(y_{t_1},\ldots,y_{t_n};\,\lambda,p)}{d\lambda}\right)^2}{\mathcal{L}(y_{t_1},\ldots,y_{t_n};\,\lambda,p)}\,.$$

- By exploiting **Chebyshev's inequality**, we have

$$\begin{aligned} \Pr\left(E[Z]-12\sqrt{Var(Z)} \le Z \le E[Z]+12\sqrt{Var(Z)}\right) \;&\geq\; 1-\frac{1}{12^2} \\ &=\; 99.3\%\,. \end{aligned}$$

Theoretical Result

Proposition (Bean, Eshragh and Ross; 2015)

For a POSBP with n observations and time horizon τ, the **optimal observation time** *for the last observation, that is t_n^*, is equal to τ.*

Theoretical Result

Proposition (Bean, Eshragh and Ross; 2015)

For a POSBP with n observations and time horizon τ, the **optimal observation time** *for the last observation, that is t_n^*, is equal to $\boldsymbol{\tau}$.*

Proposition (Bean, Eshragh and Ross; 2015)

If $t_1^*, \ldots, t_n^*$ are optimal observation times for a POSBP with parameters (λ, p) and time-horizon τ, then $\frac{\mathbf{t_1^*}}{\boldsymbol{\tau}}, \ldots, \frac{\mathbf{t_n^*}}{\boldsymbol{\tau}}$ are **optimal observation times** *for a POSBP with parameters $(\boldsymbol{\lambda\tau}, p)$ and time-horizon $\mathbf{1}$.*

Theoretical Result

Proposition (Bean, Eshragh and Ross; 2015)

For a POSBP with n observations and time horizon τ, the **optimal observation time** *for the last observation, that is t_n^*, is equal to $\boldsymbol{\tau}$.*

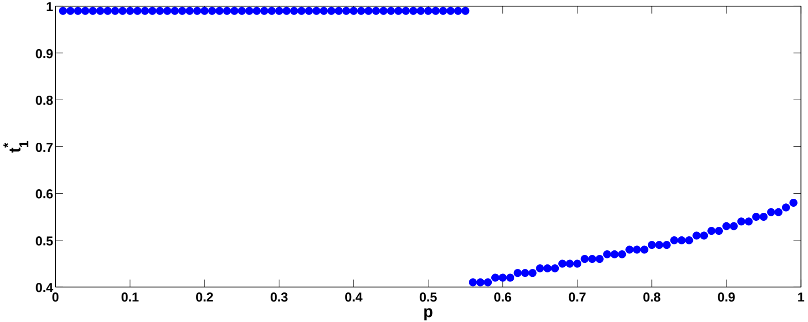
Proposition (Bean, Eshragh and Ross; 2015)

If $t_1^*, \ldots, t_n^*$ are optimal observation times for a POSBP with parameters (λ, p) and time-horizon τ, then $\frac{\mathbf{t_1^*}}{\boldsymbol{\tau}}, \ldots, \frac{\mathbf{t_n^*}}{\boldsymbol{\tau}}$ are **optimal observation times** *for a POSBP with parameters $(\boldsymbol{\lambda\tau}, p)$ and time-horizon $\mathbf{1}$.*

- Henceforth, **without loss of generality**, we assume that $\boldsymbol{\tau} = \mathbf{1}\ (= \mathbf{t_n^*})$.

Results for $\lambda = 2$, $n = 2$ and $t_2^* = \tau = 1$

- Optimal observation time t_1^* vs. p



The Chain Rule

- The likelihood function

$$\mathcal{L}(y_{t_1}, y_{t_2};\ \lambda, p) = \Pr(Y_{t_2} = y_{t_2} | Y_{t_1} = y_{t_1},\ \lambda) \Pr(Y_{t_1} = y_{t_1} | \lambda) .$$

The Chain Rule

- The likelihood function

$$\mathcal{L}(y_{t_1}, y_{t_2};\ \lambda, p) = \Pr(Y_{t_2} = y_{t_2} | Y_{t_1} = y_{t_1},\ \lambda) \Pr(Y_{t_1} = y_{t_1} | \lambda) .$$

- Accordingly,

$$\begin{aligned}\log\left(\mathcal{L}(y_{t_1}, y_{t_2};\ \lambda, p)\right) = &\log\left(\Pr(Y_{t_2} = y_{t_2} | Y_{t_1} = y_{t_1},\ \lambda)\right) \\ &+ \log\left(\Pr(Y_{t_1} = y_{t_1} | \lambda)\right) .\end{aligned}$$

The Chain Rule

- The likelihood function

$$\mathcal{L}(y_{t_1}, y_{t_2};\ \lambda, p) = \Pr(Y_{t_2} = y_{t_2} | Y_{t_1} = y_{t_1},\ \lambda) \Pr(Y_{t_1} = y_{t_1} | \lambda)\,.$$

- Accordingly,

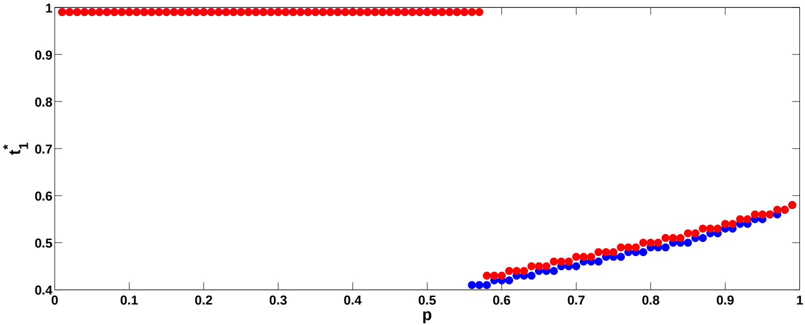
$$\begin{aligned}\log\left(\mathcal{L}(y_{t_1}, y_{t_2};\ \lambda, p)\right) = & \log\left(\Pr(Y_{t_2} = y_{t_2} | Y_{t_1} = y_{t_1},\ \lambda)\right) \\ & + \log\left(\Pr(Y_{t_1} = y_{t_1} | \lambda)\right).\end{aligned}$$

- Fisher Information:

$$\mathcal{FI}_{(Y_{t_1}, Y_{t_2})}(\lambda) \quad = \quad \mathcal{FI}_{(Y_{t_2}|Y_{t_1})}(\lambda) + \mathcal{FI}_{(Y_{t_1})}(\lambda)\,.$$

Results for $\lambda = 2$, $n = 2$ and $t_2^* = \tau = 1$

- Optimal observation time t_1^* vs. p



Experimental Mathematics Approach

- Construct the **generating function** for the likelihood function:

$$\phi(u_1, \ldots, u_n) \quad = \quad \sum_{y_{t_n}=0}^{\infty} \cdots \sum_{y_{t_1}=0}^{\infty} \mathcal{L}_{Y_n}(y_1, \ldots, y_n;\, \lambda, p) \prod_{i=1}^{n} u_i^{y_{t_i}}$$

Experimental Mathematics Approach

- Construct the **generating function** for the likelihood function:

$$
\begin{aligned}
\phi(u_1, \ldots, u_n) &= \sum_{y_{t_n}=0}^{\infty} \cdots \sum_{y_{t_1}=0}^{\infty} \mathcal{L}_{Y_n}(y_1, \ldots, y_n;\ \lambda, p) \prod_{i=1}^{n} u_i^{y_{t_i}} \\
&= \frac{P(u_1, \ldots, u_n)}{Q(u_1, \ldots, u_n)}
\end{aligned}
$$

Experimental Mathematics Approach

- Construct the **generating function** for the likelihood function:

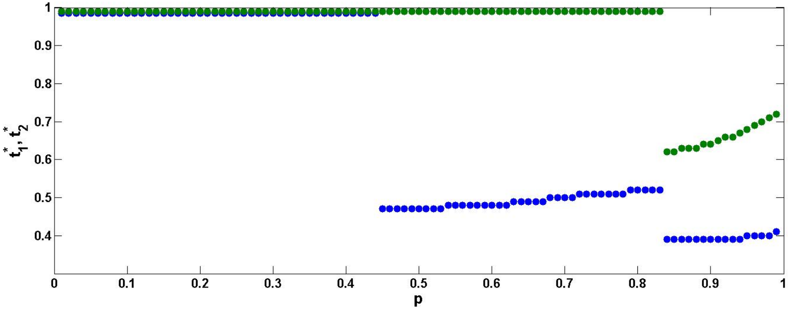
$$\begin{aligned}\phi(u_1,\ldots,u_n) &= \sum_{y_{t_n}=0}^{\infty}\cdots\sum_{y_{t_1}=0}^{\infty}\mathcal{L}_{Y_n}(y_1,\ldots,y_n;\,\lambda,p)\prod_{i=1}^{n}u_i^{y_{t_i}}\\ &= \frac{P(u_1,\ldots,u_n)}{Q(u_1,\ldots,u_n)}\end{aligned}$$

- Once the **polynomial functions** P and Q are found, one can construct a **recursive equation** for the likelihood function by equating

$$Q(u_1,\ldots,u_n)\sum_{y_n=0}^{\infty}\cdots\sum_{y_1=0}^{\infty}\mathcal{L}_{Y_n}(y_1,\ldots,y_n;\,\lambda,p)\prod_{i=1}^{n}u_i^{y_{t_i}} \equiv P(u_1,\ldots,u_n)\,.$$

Results for $\lambda = 2$, $n = 3$ and $t_3^* = \tau = 1$

- Optimal observation times t_1^* (blue) and t_2^* (green) vs. p



Summary

The **Fisher Information** for the **POSBP**

Summary

The **Fisher Information** for the **POSBP**

- is **analytically intractable** even when there is only one observation;

Summary

The **Fisher Information** for the **POSBP**

- is **analytically intractable** even when there is only one observation;
- could be calculated **numerically** only for $\lambda \leq 2$ and $n = 2$ in significant run-time by **truncating** the infinite sums;

Summary

The **Fisher Information** for the **POSBP**

- is **analytically intractable** even when there is only one observation;
- could be calculated **numerically** only for $\lambda \leq 2$ and $n = 2$ in significant run-time by **truncating** the infinite sums;
- was **approximated** very quickly for any value of λ and $n = 2$ by exploiting **Applied Probability** concepts;

Summary

The **Fisher Information** for the **POSBP**

- is **analytically intractable** even when there is only one observation;
- could be calculated **numerically** only for $\lambda \leq 2$ and $n = 2$ in significant run-time by **truncating** the infinite sums;
- was **approximated** very quickly for any value of λ and $n = 2$ by exploiting **Applied Probability** concepts;
- could be calculated **numerically** for any values of λ and n in significant run-time by utilizing **Experimental Mathematics** techniques;

Summary

The **Fisher Information** for the **POSBP**

- is **analytically intractable** even when there is only one observation;
- could be calculated **numerically** only for $\lambda \leq 2$ and $n = 2$ in significant run-time by **truncating** the infinite sums;
- was **approximated** very quickly for any value of λ and $n = 2$ by exploiting **Applied Probability** concepts;
- could be calculated **numerically** for any values of λ and n in significant run-time by utilizing **Experimental Mathematics** techniques; and surprisingly could **reduce** the run-time by a factor of at least **32,000**.

End

Thank you $\cdots$ Questions?