Simulation-Based Estimation

Michael W. Trosset
Department of Mathematics
College of William & Mary

trosset@math.wm.edu

http://www.math.wm.edu/~trosset/

Introduction

Stochastic simulations try to mimic stochastic phenomena observed in nature.

The natural phenomenon is modeled by a probability distribution Q. We observe what Nature draws from Q. Independently replicated observations constitute a sample $\vec{y} = \{y_1, \ldots, y_m\}$.

Analogously, a stochastic simulation is a pseudorandom mechanism for drawing observations from a specified probability distribution. Simulation usually requires specifying the values of various parameters, so we write P_θ to denote simulation with parameter values θ. Running the simulation with these values means drawing observations from P_θ. Independently replicated observations constitute a sample $\vec{x} = \{x_1, \ldots, x_n\}$.

Parameter Estimation

Suppose that we've observed an *actual sample* $\vec{y}$, drawn by Nature from Q.

Problem: Guess what parameter value causes P_θ to behave most like Q.

Strategy: Draw simulated samples from different P_θ. Choose the value(s) for which the simulated samples most closely resemble the actual sample.

Key questions:

1. How do we measure how closely two samples resemble each other?

   (For simplicity, assume that we're observing real numbers, e.g. time from first tumor detected to second tumor detected.)

2. How do we systematically choose different P_θ from which to sample?

Comparing Q and $P(\theta)$

1. Partition $(-\infty, \infty)$ into b bins: $B_1, \ldots, B_b$.

2. Let

$$q_k = Q(B_k)$$

and

$$p_k(\theta) = P_\theta(B_k).$$

3. Let

$$\Delta\left(Q, P_\theta\right) = m \sum_{k=1}^{b} \frac{\left[q_k - p_k(\theta)\right]^2}{p_k(\theta)}.$$

4. Choose θ to minimize $\Delta(Q, P_\theta)$.

Unfortunately, Q is unknown—we only know $\vec{y}$, the sample drawn from Q.

Comparing $\vec{y} \sim Q$ and $P(\theta)$

1. Partition $(-\infty, \infty)$ into b bins: $B_1, \ldots, B_b$.

2. Let

$$\widehat{q}_k = \# \left\{ y_j \in B_k \right\} / m$$

   and

$$p_k(\theta) = P_\theta(B_k).$$

3. Let

$$f(\theta) = \Delta\left(\widehat{Q}, P_\theta\right) = m \sum_{k=1}^{b} \frac{[\widehat{q}_k - p_k(\theta)]^2}{p_k(\theta)}.$$

4. Choose $\widehat{\theta}$ to minimize $f(\theta)$.

Unfortunately, most stochastic simulations are too complicated for us to determine $p_k(\theta)$. Instead, we must sample P_θ in order to estimate $p_k(\theta)$.

Comparing $\vec{y} \sim Q$ and $\vec{x} \sim P(\theta)$

1. Partition $(-\infty, \infty)$ into b bins: $B_1, \ldots, B_b$.

2. Let

$$\widehat{q}_k = \# \left\{ y_j \in B_k \right\} / m$$

   and

$$\widehat{p}_k(\theta) = \# \left\{ x_i \in B_k \right\} / n.$$

3. Let

$$\widehat{f}_n(\theta) = \Delta\left(\widehat{Q}, \widehat{P}_\theta\right) = m \sum_{k=1}^{b} \frac{\left[\widehat{q}_k - \widehat{p}_k(\theta)\right]^2}{\widehat{p}_k(\theta)}.$$

4. Choose $\widehat{\theta}$ to minimize $f(\theta)$, using $\widehat{f}_n(\theta)$ to estimate $f(\theta)$.

Notice that function evaluation is uncertain—drawing different samples from P_θ will produce different $\widehat{f}_n(\theta)$.

However, $\widehat{f}_n(\theta) \xrightarrow{P} f(\theta)$ as $n \to \infty$.

Example

Bin	Q	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
$(-\infty, 2)$	5	67	64	27	36	12	13
$(2, 3)$	6	19	16	26	20	22	23
$(3, 4)$	4	6	11	23	21	18	20
$(4, 5)$	2	4	5	10	12	21	19
$(5, \infty)$	3	4	4	14	11	27	25

θ	$\widehat{f}_n(\theta)$
θ_1	20.92
θ_2	15.73
θ_3	0.25
θ_4	2.04
θ_5	5.66
θ_6	4.29

Answer

$$\theta_1 = \theta_2 = 2$$
$$\theta_3 = \theta_4 = 3$$
$$\theta_5 = \theta_6 = 4$$

$$Q = P_\pi$$

Next: Stochastic Optimization

We want to minimize f.

Function evaluation is uncertain: because we can't compute $f(\theta)$, we must estimate $f(\theta)$ by sampling P_θ and computing $\widehat{f}_n(\theta)$.

Question: How do we minimize when function evaluation is uncertain?

Papers to be Presented

1. What is P_θ, the stochastic simulation?

2. How are samples compared?

3. How is the comparison criterion optimized?