Stochastic Simulation

Jan-Pieter Dorsman[1] & Michel Mandjes[1,2,3]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]CWI, Amsterdam
[3]Eurandom, Eindhoven

University of Amsterdam,
Fall, 2017

Chapter VII
Derivative Estimation

The idea behind derivative estimation

We have seen right now how to go about simulating a value $z = \mathbb{E}[Z(\boldsymbol{\theta})]$, where Z is a random variable which depends on the parameters $\theta_1, \theta_2, \ldots$.

For purposes of sensitivity analysis, we are interested in the gradient

$$\nabla z = \left(\frac{\partial}{\partial \theta_1} z \quad \frac{\partial}{\partial \theta_2} z \quad \cdots \right).$$

We now adress the problem of how to estimate this gradient by simulation, i.e. derivative estimation.

Why?

There are numerous reasons why we'd be interested in estimating the gradient:

- To identify the most important system parameter.
- To assess the effect of a small parameter change.
- In optimization, to find the best system parameter $\boldsymbol{\theta}$ requires evaluation of the gradient (think of Newton-Raphson).
- To find gradients that are of intrinsic interest, such as the Greeks in option pricing.

Why?

There are numerous reasons why we'd be interested in estimating the gradient:

- To identify the most important system parameter.
- To assess the effect of a small parameter change.
- In optimization, to find the best system parameter θ requires evaluation of the gradient (think of Newton-Raphson).
- To find gradients that are of intrinsic interest, such as the Greeks in option pricing.

From a theoretical point of view, there is hardly any difference between the one-dimensional and the multi-dimensional case, as gradients and Hessians are computed componentwise. Therefore, we focus on the one-dimensional setting.

Derivative estimation

We discuss three methods.

1. Finite differences (FD)
2. Infinitesimal perturbation analysis (IPA)
3. Likelihood ratio method (LR)

Let's say that $Z = h(X)$. Important differences:

- For FD and IPA, we only assume the function h to depend on θ (structural dependence).
- For LR, we only assume the distribution of X to depend on θ (distributional dependence).

Derivative estimation

Our goal: find a random variable $D(\theta)$ such that

- in case dependence is structural,

$$\mathbb{E}\left[D(\theta)\right] = z'(\theta) = \frac{d}{d\theta}\mathbb{E}\left[h_\theta(X)\right].$$

- in case dependence is distributional,

$$\mathbb{E}\left[D(\theta)\right] = z'(\theta) = \frac{d}{d\theta}\mathbb{E}_\theta\left[h(X)\right].$$

Then, if we have found an unbiased estimator, we could e.g. use crude Monte Carlo (or other methods) to estimate $z'(\theta)$:

1. Generate R copies $D_1(\theta), \dots, D_R(\theta)$.
2. Compute $\frac{1}{R}\sum_{i=1}^{R} D_i(\theta)$.

Chapter VII.1
The finite differences method

Finite differences

Suppose that for each θ, we can generate an r.v. $Z(\theta)$ with expectation $z(\theta)$.

Starting point is the definition of a derivative:

$$f'(\theta) = \lim_{h \to 0} \frac{f(\theta + h) - f(\theta)}{h} = \lim_{h \to 0} \frac{f(\theta + \frac{h}{2}) - f(\theta - \frac{h}{2})}{h}$$

Finite differences

Suppose that for each θ, we can generate an r.v. $Z(\theta)$ with expectation $z(\theta)$.

Starting point is the definition of a derivative:

$$f'(\theta) = \lim_{h \to 0} \frac{f(\theta + h) - f(\theta)}{h} = \lim_{h \to 0} \frac{f(\theta + \frac{h}{2}) - f(\theta - \frac{h}{2})}{h}$$

This suggests two possible derivative estimators:

$$\tilde{D}(\theta) = \frac{Z(\theta + h) - Z(\theta)}{h}$$

or

$$D(\theta) = \frac{Z(\theta + \frac{h}{2}) - Z(\theta - \frac{h}{2})}{h}.$$

Finite differences

Suppose that for each θ, we can generate an r.v. $Z(\theta)$ with expectation $z(\theta)$.

Starting point is the definition of a derivative:

$$f'(\theta) = \lim_{h \to 0} \frac{f(\theta + h) - f(\theta)}{h} = \lim_{h \to 0} \frac{f(\theta + \frac{h}{2}) - f(\theta - \frac{h}{2})}{h}$$

This suggests two possible derivative estimators:

$$\tilde{D}(\theta) = \frac{Z(\theta + h) - Z(\theta)}{h}$$

or

$$D(\theta) = \frac{Z(\theta + \frac{h}{2}) - Z(\theta - \frac{h}{2})}{h}.$$

Q: Should we use the forward difference estimator $\tilde{D}(\theta)$ or the central difference estimator $D(\theta)$?

Finite differences

Suppose that for each θ, we can generate an r.v. $Z(\theta)$ with expectation $z(\theta)$.

Starting point is the definition of a derivative:

$$f'(\theta) = \lim_{h\to 0} \frac{f(\theta+h) - f(\theta)}{h} = \lim_{h\to 0} \frac{f(\theta+\frac{h}{2}) - f(\theta-\frac{h}{2})}{h}$$

This suggests two possible derivative estimators:

$$\tilde{D}(\theta) = \frac{Z(\theta+h) - Z(\theta)}{h}$$

or

$$D(\theta) = \frac{Z(\theta+\frac{h}{2}) - Z(\theta-\frac{h}{2})}{h}.$$

Q: Should we use the forward difference estimator $\tilde{D}(\theta)$ or the central difference estimator $D(\theta)$?

A: As always, check biasedness first.

Finite differences

To check biasedness, apply Taylor series about θ.

$$\begin{aligned}\mathbb{E}\left[\tilde{D}(\theta)\right] &= \frac{z(\theta+h)-z(\theta)}{h} \\ &= \frac{1}{h}\left(z'(\theta)h + z''(\theta)\frac{h^2}{2} + z'''(\theta)\frac{h^3}{6} + \dots\right)\end{aligned}$$

Finite differences

To check biasedness, apply Taylor series about θ.

$$\mathbb{E}\left[\tilde{D}(\theta)\right] = \frac{z(\theta+h) - z(\theta)}{h}$$
$$= \frac{1}{h}\left(z'(\theta)h + z''(\theta)\frac{h^2}{2} + z'''(\theta)\frac{h^3}{6} + \dots\right)$$

and

$$\mathbb{E}\left[D(\theta)\right] = \frac{z(\theta+\frac{h}{2}) - z(\theta-\frac{h}{2})}{h}$$
$$= \frac{1}{h}\left(z'(\theta)h + z'''(\theta)\frac{h^3}{24} + \dots\right)$$

Conclusion?

Finite differences

To check biasedness, apply Taylor series about θ.

$$\mathbb{E}\left[\tilde{D}(\theta)\right] = \frac{z(\theta + h) - z(\theta)}{h}$$
$$= \frac{1}{h}\left(z'(\theta)h + z''(\theta)\frac{h^2}{2} + z'''(\theta)\frac{h^3}{6} + \dots\right)$$

and

$$\mathbb{E}\left[D(\theta)\right] = \frac{z(\theta + \frac{h}{2}) - z(\theta - \frac{h}{2})}{h}$$
$$= \frac{1}{h}\left(z'(\theta)h + z'''(\theta)\frac{h^3}{24} + \dots\right)$$

Conclusion? The bias of $D(\theta)$ is an order of magnitude lower, so let's go with that estimator.

Q: How to choose h?

Finite differences

To check biasedness, apply Taylor series about θ.

$$\begin{aligned}\mathbb{E}\left[\tilde{D}(\theta)\right] &= \frac{z(\theta+h)-z(\theta)}{h} \\ &= \frac{1}{h}\left(z'(\theta)h + z''(\theta)\frac{h^2}{2} + z'''(\theta)\frac{h^3}{6} + \dots\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[D(\theta)\right] &= \frac{z(\theta+\frac{h}{2})-z(\theta-\frac{h}{2})}{h} \\ &= \frac{1}{h}\left(z'(\theta)h + z'''(\theta)\frac{h^3}{24} + \dots\right)\end{aligned}$$

Conclusion? The bias of $D(\theta)$ is an order of magnitude lower, so let's go with that estimator.

Q: How to choose h?

A: That's a good question!

Finite differences

How to choose h?

- On one hand, don't pick h too large, because it induces bias. The bias will vanish as $h \to 0$.
- On other hand, don't pick h too small, because it increases variance.

Finite differences

How to choose h?

- On one hand, don't pick h too large, because it induces bias. The bias will vanish as $h \to 0$.
- On other hand, don't pick h too small, because it increases variance.

When estimating $z(\theta + \frac{h}{2})$ and $z(\theta - \frac{h}{2})$, each independently with R samples, then the mean squared error of the estimator is optimised when choosing

$$h = \frac{1}{R^{1/6}} \frac{(576 Var(Z(\theta)))^{1/6}}{|z'''(\theta)|^{1/3}},$$

in which case the root of the mean squared error is of order $R^{-1/3}$. See book for the proof.

Of course, this result is rather academic. When estimating $z'(\theta)$, one probably doesn't know $z'''(\theta)$, but dependence on R is interesting.

Finite differences

Tips and tricks:

- One can use common random numbers to reduce variance!
- For instance, suppose that the dependence is distributional: $Z = g(X)$, where θ is a parameter of the pdf of X.
- Generate independent uniform samples $U_1, \ldots, U_R$.
- Compute

$$Z_i^{(+)} = g\left(F^{\leftarrow}_{\theta+\frac{h}{2}}(U_i)\right); \qquad Z_i^{(-)} = g\left(F^{\leftarrow}_{\theta-\frac{h}{2}}(U_i)\right);$$

for all $i = 1, \ldots, R$.

Finite differences

Tips and tricks:

- One can use common random numbers to reduce variance!
- For instance, suppose that the dependence is distributional: $Z = g(X)$, where θ is a parameter of the pdf of X.
- Generate independent uniform samples $U_1, \ldots, U_R$.
- Compute

$$Z_i^{(+)} = g\left(F^{\leftarrow}_{\theta+\frac{h}{2}}(U_i)\right); \qquad Z_i^{(-)} = g\left(F^{\leftarrow}_{\theta-\frac{h}{2}}(U_i)\right);$$

  for all $i = 1, \ldots, R$.
- Finite difference estimator for $z'(\theta)$ now is

$$\frac{1}{hR}\sum_{i=1}^{R}(Z_i^{(+)} - Z_i^{(-)}).$$

Finite differences

Example.

- Suppose that $Z = h(X)$, where X is $\exp(\theta)$ distributed and $g(x) = x^p$ with $p \in \mathbb{N}$.
- Note that $z(\theta) = \mathbb{E}[X^p]$ is the p-th moment of the $\exp(\theta)$ distribution, i.e. $z(\theta) = p! \left(\frac{1}{\theta}\right)^p$, but let's suppose we unknowingly wish to estimate $z'(\theta)$.

Finite differences

Example.

- Suppose that $Z = h(X)$, where X is $\exp(\theta)$ distributed and $g(x) = x^p$ with $p \in \mathbb{N}$.
- Note that $z(\theta) = \mathbb{E}[X^p]$ is the p-th moment of the $\exp(\theta)$ distribution, i.e. $z(\theta) = p!\left(\frac{1}{\theta}\right)^p$, but let's suppose we unknowingly wish to estimate $z'(\theta)$.
- Recall that the quantile function of X is given by $F^{\leftarrow}(x) = \frac{-\log(1-x)}{\theta}$.
- This leads to

$$D(\theta) = \frac{1}{h}\left(\left(\frac{1}{\theta + \frac{h}{2}}\right)^p - \left(\frac{1}{\theta - \frac{h}{2}}\right)^p\right)(-log(U))^p,$$

where U is standard uniformly distributed.

Finite differences

Common random variables versus independent sampling.

It can be seen that using this estimator,

$$\mathrm{Var}\left[D(\theta)\right] = \frac{1}{h^2}\left(\left(\frac{1}{\theta+\frac{h}{2}}\right)^{p} - \left(\frac{1}{\theta-\frac{h}{2}}\right)^{p}\right)^{2}\left((2p)! - (p!)^2\right).$$

However, in case we wouldn't have used common random numbers but sampled independently, we would have faced

$$\mathrm{Var}\left[D(\theta)\right] = \frac{1}{h^2}\left(\left(\frac{1}{\theta+\frac{h}{2}}\right)^{2p} + \left(\frac{1}{\theta-\frac{h}{2}}\right)^{2p}\right)\left((2p)! - (p!)^2\right),$$

which is larger!

Finite differences

The book also discusses higher-order finite-difference approximations, such as the second-order approximation

$$\frac{-Z(\theta + 2h) + 4Z(\theta + h) - 3Z(\theta)}{2h}$$

and much higher order. These k-order estimators typically have a smaller bias (roughly of order h^k), but

- these estimators require approximations at $k + 1$ points of $z(\cdot)$,
- the weights involved grow larger in absolute value as k gets high, doing the variance of the estimator no good.

Finite differences

The book also discusses higher-order finite-difference approximations, such as the second-order approximation

$$\frac{-Z(\theta+2h)+4Z(\theta+h)-3Z(\theta)}{2h}$$

and much higher order. These k-order estimators typically have a smaller bias (roughly of order h^k), but

- these estimators require approximations at $k+1$ points of $z(\cdot)$,
- the weights involved grow larger in absolute value as k gets high, doing the variance of the estimator no good.

Therefore, higher-order approximations are hardly used in practice.

Chapter VII.2
Infinitesimal Perturbation Analysis

Infinitesimal Perturbation Analysis

Suppose that $Z = h_\theta(X)$ (structural dependence), and that we want to know $\frac{d}{d\theta}\mathbb{E}[Z]$.

Let $D(\theta) := \frac{d}{d\theta}h_\theta(X)$. Then, IPA is based on the following assumption:

$$\frac{d}{d\theta}\mathbb{E}[h_\theta(X)] = \mathbb{E}\left[\frac{d}{d\theta}h_\theta(X)\right] = \mathbb{E}[D(\theta)].$$

Infinitesimal Perturbation Analysis

Suppose that $Z = h_\theta(X)$ (structural dependence), and that we want to know $\frac{d}{d\theta}\mathbb{E}\left[Z\right]$.

Let $D(\theta) := \frac{d}{d\theta}h_\theta(X)$. Then, IPA is based on the following assumption:

$$\frac{d}{d\theta}\mathbb{E}\left[h_\theta(X)\right] = \mathbb{E}\left[\frac{d}{d\theta}h_\theta(X)\right] = \mathbb{E}\left[D(\theta)\right].$$

The IPA-method simply implies the crude Monte Carlo simulation of $D(\theta)$:

1. Sample R copies of $D(\theta)$.
2. Create confidence intervals based on these copies using regular methods and techniques.

Infinitesimal Perturbation Analysis

Same example as before:

- Suppose that $Z = g(X)$, where X is $\exp(\theta)$ distributed and $g(x) = x^p$ with $p \in \mathbb{N}$.
- Z has same distribution as $h_\theta(U)$, where $h_\theta(x) = (-\frac{\log(x)}{\theta})^p$, and U is standard uniform.

Infinitesimal Perturbation Analysis

Same example as before:

- Suppose that $Z = g(X)$, where X is $\exp(\theta)$ distributed and $g(x) = x^p$ with $p \in \mathbb{N}$.
- Z has same distribution as $h_\theta(U)$, where $h_\theta(x) = (-\frac{\log(x)}{\theta})^p$, and U is standard uniform.
- We have

$$\frac{d}{d\theta} h_\theta(x) = -(-\log(x))^p p \left(\frac{1}{\theta}\right)^{p-1} \frac{1}{\theta^2} = \frac{-p h_\theta(x)}{\theta}.$$

- This results in

$$D(\theta) = \frac{-p h_\theta(U)}{\theta}$$

- Thus: create replicates of $h_\theta(U)$ using known methods, and then multiply each with $\frac{-p}{\theta}$. These are the resulting replicates for $D(\theta)$.

Infinitesimal Perturbation Analysis

Same example as before:

- Suppose that $Z = g(X)$, where X is $\exp(\theta)$ distributed and $g(x) = x^p$ with $p \in \mathbb{N}$.
- Z has same distribution as $h_\theta(U)$, where $h_\theta(x) = (-\frac{\log(x)}{\theta})^p$, and U is standard uniform.
- We have

$$\frac{d}{d\theta} h_\theta(x) = -(-\log(x))^p p \left(\frac{1}{\theta}\right)^{p-1} \frac{1}{\theta^2} = \frac{-p h_\theta(x)}{\theta}.$$

- This results in

$$D(\theta) = \frac{-p h_\theta(U)}{\theta}$$

- Thus: create replicates of $h_\theta(U)$ using known methods, and then multiply each with $\frac{-p}{\theta}$. These are the resulting replicates for $D(\theta)$.

We can get estimates for different values of θ with negligible additional effort.

Infinitesimal Perturbation Analysis

Q: Are we done now?

Infinitesimal Perturbation Analysis

Q: Are we done now?

A: No, who says we can interchange expectation and derivative like that?

Infinitesimal Perturbation Analysis

Q: Are we done now?

A: No, who says we can interchange expectation and derivative like that?

Proposition: Assume that $Z(\theta)$ is a.s. differentiable at θ_0 and that a.s. $Z(\theta)$ satisfies the Lipschitz condition

$$|Z(\theta_1) - Z(\theta_2)| \leq |\theta_1 - \theta_2| M$$

for θ_1, θ_2 in a nonrandom neighborhood of θ_0, where $\mathbb{E}[M] < \infty$. Then,

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)]\bigg|_{\theta=\theta_0} = \mathbb{E}\left[\frac{d}{d\theta}Z(\theta)\right]\bigg|_{\theta=\theta_0}.$$

Infinitesimal Perturbation Analysis

Q: Are we done now?

A: No, who says we can interchange expectation and derivative like that?

Proposition: Assume that $Z(\theta)$ is a.s. differentiable at θ_0 and that a.s. $Z(\theta)$ satisfies the Lipschitz condition

$$|Z(\theta_1) - Z(\theta_2)| \leq |\theta_1 - \theta_2| M$$

for θ_1, θ_2 in a nonrandom neighborhood of θ_0, where $\mathbb{E}[M] < \infty$. Then,

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)]\bigg|_{\theta=\theta_0} = \mathbb{E}\left[\frac{d}{d\theta}Z(\theta)\right]\bigg|_{\theta=\theta_0}.$$

Proof: Use of dominated convergence theorem.

Infinitesimal Perturbation Analysis

Example of when the interchangeability assumption fails.

- Let X_1, X_2 be two independent r.v.s, and let $z = \mathbb{P}(X_1 < \theta X_2)$ be the expectation of $Z = \mathbb{1}_{\{X_1 < \theta X_2\}}$.
- Now, $D(\theta) = \frac{d}{d\theta} Z = 0$ a.s. for any value of θ. But $\frac{d}{d\theta} z$ surely is not zero in general for any value of θ.

The proposition indeed does not apply. Let's say that we want to know $\frac{d}{d\theta} z$ in the point $\theta_0 = 1$.

Then, we have for $\theta \in [1, 1 + \epsilon)$ that

$$Z(\theta) - Z(\theta_0) = \mathbb{1}_{\{\theta > X_1/X_2\}} - \mathbb{1}_{\{1 > X_1/X_2\}} = \mathbb{1}_{\{1 \le X_1/X_2 < \theta\}}.$$

$$Z(\theta) - Z(\theta_0) = \mathbb{1}_{\{\theta > X_1/X_2\}} - \mathbb{1}_{\{1 > X_1/X_2\}} = \mathbb{1}_{\{1 \le X_1/X_2 < \theta\}}.$$

Let's say that the density of X_1/X_2 is at least $A > 0$ in the interval $[1, 1+\epsilon)$. Then, this equals one with probability at least $A(\theta - 1)$. Now, in order for

$$|Z(\theta) - Z(1)| \le (\theta - 1)M$$

to hold, it must thus hold that $\mathbb{P}\left((\theta - 1)M \ge 1\right) \ge A(\theta - 1)$ for $\theta \in [1, 1+\epsilon)$.

$$Z(\theta) - Z(\theta_0) = \mathbb{1}_{\{\theta > X_1/X_2\}} - \mathbb{1}_{\{1 > X_1/X_2\}} = \mathbb{1}_{\{1 \le X_1/X_2 < \theta\}}.$$

Let's say that the density of X_1/X_2 is at least $A > 0$ in the interval $[1, 1+\epsilon)$. Then, this equals one with probability at least $A(\theta - 1)$. Now, in order for

$$|Z(\theta) - Z(1)| \le (\theta - 1)M$$

to hold, it must thus hold that $\mathbb{P}\left((\theta-1)M \ge 1\right) \ge A(\theta - 1)$ for $\theta \in [1, 1+\epsilon)$. Or, $\mathbb{P}\left(M \ge x\right) \ge \frac{A}{x}$ for all $x \in (\epsilon^{-1}, \infty)$, so that

$$\mathbb{E}\left[M\right] = \int_{x=0}^{\infty} \mathbb{P}\left(M > x\right) dx \ge \int_{x=\epsilon^{-1}}^{\infty} \mathbb{P}\left(M > x\right) dx$$
$$\ge A \int_{x=\epsilon^{-1}}^{\infty} \frac{1}{x} dx = \infty.$$

Violation of assumption!

Infinitesimal Perturbation Analysis

This effect also occurs often when step functions are involved. Example of when it does work:

- Let $Z = \max(X_1, \theta X_2)$, where Z_1 and Z_2 are two independent r.v.s.
- Then, $D(\theta) = X_2 \mathbb{1}_{\{X_1 < \theta X_2\}}$ is a valid estimator for $\frac{d}{d\theta}\mathbb{E}\left[Z(\theta)\right]$.

Justification:

$$|Z(\theta_1) - Z(\theta_2)| \leq |\theta_1 - \theta_2||X_2|.$$

Hence, take $|X_2| = M$. The interchange of derivative and expectation is now justified, unless $\mathbb{E}\left[|X_2|\right] = \infty$.

Chapter VII.3
The likelihood ratio method

The likelihood ratio method

For this method to work, we assume the dependence on θ to be distributional, i.e.

$$Z(\theta) = h(X), \qquad z(\theta) = \mathbb{E}_\theta\left[h(X)\right]$$

where X has a density $f_\theta(x)$, and we wish to estimate $z'(\theta)$ in the point θ_0. The likelihood ratio method is reminiscent of importance sampling. Let $L(\theta, x) = \frac{f_\theta(x)}{f_{\theta_0}(x)}$.

The likelihood ratio method

For this method to work, we assume the dependence on θ to be distributional, i.e.

$$Z(\theta) = h(X), \qquad z(\theta) = \mathbb{E}_\theta\left[h(X)\right]$$

where X has a density $f_\theta(x)$, and we wish to estimate $z'(\theta)$ in the point θ_0. The likelihood ratio method is reminiscent of importance sampling. Let $L(\theta, x) = \frac{f_\theta(x)}{f_{\theta_0}(x)}$. Then,

$$\begin{aligned} z(\theta) &= \int h(x) f_\theta(x) dx = \int h(x) L(\theta, x) f_{\theta_0}(x) dx \\ &= \mathbb{E}_{\theta_0}\left[h(X) L(\theta, X)\right], \end{aligned}$$

where $\mathbb{E}_{\theta_0}\left[\cdot\right]$ indicates expectation with respect to the density f_{θ_0}.

The likelihood ratio method

We have

$$z(\theta) = \mathbb{E}_{\theta_0}\left[h(X)L(\theta, X)\right].$$

The likelihood ratio method

We have

$$z(\theta) = \mathbb{E}_{\theta_0}\left[h(X)L(\theta, X)\right].$$

This suggests that

$$z'(\theta) = \mathbb{E}_{\theta_0}\left[h(X)L'(\theta, X)\right],$$

where $L'(\theta, x) = \frac{d}{d\theta}L(\theta, X)$. We can write

$$z'(\theta_0) = \mathbb{E}_{\theta_0}\left[h(X)S_{\theta_0}(X)\right],$$

where

$$S_{\theta_0}(x) = \left.\frac{f'_\theta(x)}{f_\theta(x)}\right|_{\theta=\theta_0}.$$

We will refer to $S_{\theta_0}(x)$ as the score function evaluated at $\theta = \theta_0$.

The likelihood ratio method

All this suggests the unbiased estimator for $z'(\theta_0)$:

$$D(\theta_0) = h(X)S_{\theta_0}(X)$$

Hence, we have the following estimation method:

1. Generate R i.i.d samples of X, where X has density f_{θ_0}.
2. Compute $\frac{1}{R}\sum_{i=1}^{R} h(X_i)S_{\theta_0}(X_i)$.

Confidence intervals can be computed by the known techniques.

The likelihood ratio method

This discussion was for a single random variable X. In this case,

$$S_\theta(x) = \frac{f'_\theta(x)}{f_\theta(x)} = \frac{d}{d\theta} \log f_\theta(x).$$

When $Z = h(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ have a joint density $f(\boldsymbol{x})$, the same procedure can be used. The only difference is that

$$S_\theta(\boldsymbol{x}) = \frac{d}{d\theta} \log f_\theta(\boldsymbol{x}).$$

When the X_i are independent, this leads to

$$\begin{aligned} S_\theta(\boldsymbol{x}) &= \frac{d}{d\theta} \log f_\theta(\boldsymbol{x}) = \frac{d}{d\theta} \log f_\theta^{(1)}(x_1) \cdots f_\theta^{(n)}(x_n) \\ &= \sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta^{(i)}(x_i) = \sum_{i=1}^{n} S_\theta^{(i)}(x_i). \end{aligned}$$

This is the additive property of the score function.

The likelihood ratio method

Example: $Z = h(X)$, where $h(x) = x^p$ and X is exponentially(θ) distributed.

- The corresponding score function is

$$S_\theta(x) = \frac{d}{d\theta} \log f_\theta(x) = \frac{d}{d\theta} \log(\theta) - \theta x = \frac{1}{\theta} - x$$

The likelihood ratio method

Example: $Z = h(X)$, where $h(x) = x^p$ and X is exponentially(θ) distributed.

- The corresponding score function is

$$S_\theta(x) = \frac{d}{d\theta} \log f_\theta(x) = \frac{d}{d\theta} \log(\theta) - \theta x = \frac{1}{\theta} - x$$

- Hence,

$$h(x) S_\theta(x) = \frac{x^p}{\theta} - x^{p+1}.$$

The likelihood ratio method

Example: $Z = h(X)$, where $h(x) = x^p$ and X is exponentially(θ) distributed.

- The corresponding score function is

$$S_\theta(x) = \frac{d}{d\theta} \log f_\theta(x) = \frac{d}{d\theta} \log(\theta) - \theta x = \frac{1}{\theta} - x$$

- Hence,

$$h(x) S_\theta(x) = \frac{x^p}{\theta} - x^{p+1}.$$

- So, to sample the derivative of z at $\theta = \theta_0$, we generate $X_1, \ldots, X_R$ from the $\exp(\theta_0)$ distribution, and compute

$$\frac{1}{R} \sum_{i=1}^{R} \left(\frac{X_i^p}{\theta_0} - X_i^{p+1} \right)$$

The likelihood ratio method

Example: $C = \sum_{i=1}^{N} V_i$, where N is Poisson(λ) and the V_i are i.i.d. with density $f_\theta(x)$. Assume that we are interested in $z = \mathbb{P}\left(C > x\right) = \mathbb{E}\left[\mathbb{1}_{\{C>x\}}\right]$.

- When estimating $\frac{d}{d\lambda} z|_{\lambda=\lambda_0}$, the appropriate score function is

$$S_\lambda(n) = \frac{d}{d\lambda}\log(e^{-\lambda}\frac{\lambda^n}{n!}) = \frac{n}{\lambda} - 1.$$

The likelihood ratio method

Example: $C = \sum_{i=1}^{N} V_i$, where N is Poisson(λ) and the V_i are i.i.d. with density $f_\theta(x)$. Assume that we are interested in $z = \mathbb{P}\left(C > x\right) = \mathbb{E}\left[\mathbb{1}_{\{C>x\}}\right]$.

- When estimating $\frac{d}{d\lambda} z|_{\lambda=\lambda_0}$, the appropriate score function is

$$S_\lambda(n) = \frac{d}{d\lambda} \log(e^{-\lambda} \frac{\lambda^n}{n!}) = \frac{n}{\lambda} - 1.$$

  Thus, the LR estimator is $\mathbb{1}_{\{\sum_{i=1}^{N} V_i > x\}} \left(\frac{N}{\lambda_0} - 1\right)$, where N is Poisson(λ_0) distributed and the V_i are i.i.d. with density f_θ.

The likelihood ratio method

Example: $C = \sum_{i=1}^{N} V_i$, where N is Poisson(λ) and the V_i are i.i.d. with density $f_\theta(x)$. Assume that we are interested in $z = \mathbb{P}\left(C > x\right) = \mathbb{E}\left[\mathbb{1}_{\{C>x\}}\right]$.

- When estimating $\frac{d}{d\lambda} z|_{\lambda=\lambda_0}$, the appropriate score function is

$$S_\lambda(n) = \frac{d}{d\lambda} \log(e^{-\lambda}\frac{\lambda^n}{n!}) = \frac{n}{\lambda} - 1.$$

  Thus, the LR estimator is $\mathbb{1}_{\{\sum_{i=1}^{N} V_i > x\}}\left(\frac{N}{\lambda_0} - 1\right)$, where N is Poisson(λ_0) distributed and the V_i are i.i.d. with density f_θ.
- When estimating $\frac{d}{d\theta} z|_{\theta=\theta_0}$, the appropriate score function is

$$S_\theta(v_1, \ldots, v_n) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta(v_i).$$

The likelihood ratio method

Example: $C = \sum_{i=1}^{N} V_i$, where N is Poisson(λ) and the V_i are i.i.d. with density $f_\theta(x)$. Assume that we are interested in $z = \mathbb{P}(C > x) = \mathbb{E}\left[\mathbb{1}_{\{C>x\}}\right]$.

- When estimating $\frac{d}{d\lambda} z|_{\lambda=\lambda_0}$, the appropriate score function is

$$S_\lambda(n) = \frac{d}{d\lambda} \log(e^{-\lambda}\frac{\lambda^n}{n!}) = \frac{n}{\lambda} - 1.$$

  Thus, the LR estimator is $\mathbb{1}_{\{\sum_{i=1}^{N} V_i > x\}} \left(\frac{N}{\lambda_0} - 1\right)$, where N is Poisson(λ_0) distributed and the V_i are i.i.d. with density f_θ.
- When estimating $\frac{d}{d\theta} z|_{\theta=\theta_0}$, the appropriate score function is

$$S_\theta(v_1, \ldots, v_n) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta(v_i).$$

  Then, an estimator would be $\mathbb{1}_{\{\sum_{i=1}^{N} V_i > x\}} S_{\theta_0}(V_1, \ldots, V_N)$, where N is Poisson(λ) distributed and the V_i are i.i.d. with density f_{θ_0}.

The likelihood ratio method

We forgot a minor detail in this discussion....

The likelihood ratio method

We forgot a minor detail in this discussion....

Who says derivative and expectation can be interchanged when moving from

$$z(\theta) = \mathbb{E}_{\theta_0}\left[h(X)L(\theta, X)\right]$$

to

$$z'(\theta) = \mathbb{E}_{\theta_0}\left[h(X)L'(\theta, X)\right]?$$

This again uses a dominated convergence argument (see Proposition VII.3.5), but generally, it often works out when $L(\theta, X)$ has no discontinuities depending on θ.

The likelihood ratio method

The LR method is often done in conjunction with importance sampling, so that we can estimate $z'(\theta) = \frac{d}{d\theta}\mathbb{E}[Z] = \frac{d}{d\theta}\mathbb{E}[h(X)]$ in multiple points of θ in one go.

Suppose that X has density f_{θ_0}. We then have that

$$\begin{aligned}\mathbb{E}_{\theta_0}[h(X)S_{\theta_0}(X)] &= \int h(x)S_{\theta_0}(x)f_{\theta_0}(x)dx \\ &= \int h(x)S_{\theta_0}(x)\frac{f_{\theta_0}(x)}{f_\tau(x)}f_\tau(x)dx \\ &= \mathbb{E}_\tau\left[h(X)S_{\theta_0}(X)\frac{f_{\theta_0}(X)}{f_\tau(X)}\right].\end{aligned}$$

Conclusion: we can also sample $X_1, \ldots, X_R$ as if they have density $f_\tau(x)$, and then estimate $z'(\theta_0)$ by computing

$$\frac{1}{R}\sum_{i=1}^{R} h(X_i)S_{\theta_0}(X_i)\frac{f_{\theta_0}(X_i)}{f_\tau(X_i)}.$$

We only have to sample the X_i-values once!

Discussion on the three different methods

Each of its methods has its drawbacks and advantages.

- FD is easiest to grasp, and virtually always works, but yields biased estimators. IPA and LR are unbiased.
- IPA often yields the smallest variance, but cannot always be applied.
- LR has a broader scope than IPA.