

Rare event modelling and simulation

Andrew Wood (University of Nottingham)

2 February 2009

For: Cargo Screening Network meeting, Loughborough

Outline

- ▶ My background and research interests
- ▶ The law of rare events
- ▶ Introduction to Monte Carlo techniques
- ▶ Rare-event simulation
- ▶ Multiple hypothesis testing and FDR
- ▶ Comments and discussion

Background and research interests

- ▶ I am located in the Division of Statistics, School of Mathematical Sciences, University of Nottingham.

Research interests include:

- ▶ nonparametric inference, including bootstrap and empirical likelihood methods;
- ▶ asymptotic techniques in statistics;
- ▶ simulation techniques;
- ▶ statistical shape analysis.
- ▶ analysis of microarray-type data

Current activity in applications

- ▶ Member of the University of Nottingham's Centre for Plant Integrative Biology (CPIB).
- ▶ CPIB is a large multi-disciplinary group which, in addition to biologists and bioinformaticians, includes statisticians, applied mathematicians, computer scientists and engineers.
- ▶ I am currently involved in several CPIB projects.
- ▶ I am interested in other applications of statistics and always keen to get involved in new areas.

The law of rare events

- ▶ Suppose we repeat n independent binary trials.
- ▶ The outcome of each trial is “success” with probability p , and “failure” with probability $1 - p$.
- ▶ The distribution of the number of successes, Y say, in the n trials is known as the binomial distribution, and is written $\text{Bin}(n, p)$ for short.
- ▶ The probabilities of the $\text{Bin}(n, p)$ distribution are given by

$$\text{Prob}[Y = r] = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \quad r = 0, 1, \dots, n,$$

where $n! = n \times (n-1) \dots \times 2 \times 1$.

Law of rare events, continued

- ▶ The law of rare events, also known as the law of small numbers, may be stated as follows: when p is small (so that it is a rare event for a trial to be successful) and n is large, then the distribution of Y is approximately Poisson with mean $\lambda = np$.
- ▶ In mathematical terms, this means that the distribution of Y is approximately given by

$$\text{Prob}[Y = r] = \frac{\lambda^r}{r!} e^{-\lambda}, \quad r = 0, 1, \dots$$

- ▶ This result is due to L. Bortkiewicz (1868-1931), who first applied it to data on the number of Prussian cavalry officers kicked to death by horses (presumably a rare event) each year.

Monte Carlo simulation techniques

- ▶ Basic problem: want to evaluate a quantity $\theta = E[h(X)]$.
- ▶ In the above, $E[.]$ means “expectation of”; $h(.)$ is a known function and; X is a random variable from the distribution of interest, $f(x)$ say.
- ▶ Note that θ can be expressed as an integral:

$$\int_{x \in \mathcal{X}} h(x)f(x)dx.$$

However, this integral might be difficult to evaluate directly.

- ▶ Alternative approach: generate realisations X_1, \dots, X_n from the distribution f using simulation, and then estimate θ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n h(X_i).$$

- ▶ This is the **Monte Carlo** method.

Monte Carlo, continued

- ▶ Justification: the “law of large numbers” tells us that, for large n , $\hat{\theta}$ will be close to θ , the quantity we want to calculate, with high probability.
- ▶ Note: the law of large numbers states that sample averages converge to population averages as the sample size increases.
- ▶ It can be shown that $E[\hat{\theta}] = \theta$, i.e. in statistical terminology, $\hat{\theta}$ is an unbiased estimator of θ .
- ▶ Key question 1: how to measure the performance of a Monte Carlo procedure?
- ▶ Key question 2: how to simulate random realisations of X from the distribution f ?

A measure of performance

- ▶ An important quantifier of performance is $\text{Var}(\hat{\theta})$, the variance of $\hat{\theta}$.
- ▶ Recall that variance = (standard deviation)².
- ▶ Here,

$$\text{Var}(\hat{\theta}) = V_0/n,$$

where $V_0 = \text{Var}(h(X))$ and X is a random realisation from f .

- ▶ For good performance we want $\text{Var}(\hat{\theta})$ to be as small as possible.
- ▶ Large $\text{Var}(\hat{\theta})$ implies that n , the Monte Carlo sample size, is not large enough, and/or that V_0 is problematically large .
- ▶ Many methods of variance reduction have been developed (e.g. the use of antithetic variables, control variables, Rao-Blackwellization).
- ▶ As we shall see in a moment, using variance to measure performance is *not* is not appropriate when estimating rare event probabilities.

Stochastic Simulation

- ▶ Basic problem: how to simulate X from the distribution f .
- ▶ In many problems f may be very complex and may e.g. involve an intractable normalising constant.

Stochastic Simulation

- ▶ Basic problem: how to simulate X from the distribution f .
- ▶ In many problems f may be very complex and may e.g. involve an intractable normalising constant.
- ▶ Starting point: a very long stream of Uniform[0,1] pseudo-random numbers u_1, u_2, \dots
- ▶ The u -sequences used in practice are in fact deterministic.
- ▶ However, with careful design the u -sequence will be difficult to distinguish from a truly random sequence of Uniform[0,1] variables.
- ▶ Starting with the u_1, u_2, \dots as building blocks, many techniques of simulation have been developed for different types of f , e.g. direct inversion, the rejection method, Markov Chain Monte Carlo (MCMC), sequential Monte Carlo methods.
- ▶ Usually, X may be expressed as known function g of the u -sequence u , input data z and model parameter vector β :

$$X = g(u, z, \beta).$$

Monte Carlo estimation of rare event probabilities

- ▶ This can be done within the Monte Carlo (MC) framework described above.
- ▶ Recall that the MC approach estimates the quantity $\theta = E[h(X)]$.
- ▶ Suppose $\mathcal{E} \subset \mathcal{X}$, i.e. suppose that \mathcal{E} is a subset of the sample space \mathcal{X} .
- ▶ Let $X \in \mathcal{E}$ correspond to a rare event.
- ▶ For example, if X describes the state of the system under study, let \mathcal{E} correspond to a set of “unlikely” states such the probability of any of the states in \mathcal{E} occurring is small.

Rare events, continued

- ▶ Let $I_{\mathcal{E}}(x)$ denote the *indicator function* for the set \mathcal{E} .
- ▶ In mathematical terms, this means that

$$I_{\mathcal{E}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ Define $h(x) = I_{\mathcal{E}}(x)$.
- ▶ Then

$$\theta = E[h(X)] = E[I_{\mathcal{E}}(x)] = \text{Prob}[\mathcal{E}],$$

- ▶ In the above, $\text{Prob}[\mathcal{E}]$ is the probability of the event \mathcal{E} .
- ▶ To estimate θ by Monte Carlo, simulate independent realisations X_1, \dots, X_n of X and estimate θ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n I_{\mathcal{E}}(X_i) = \frac{\#\{i : X_i \in \mathcal{E}\}}{n}$$

Rare events, continued

- ▶ In the case of rare events,

$$\text{Var}(\hat{\theta}) = \frac{p(1-p)}{n}.$$

- ▶ When p is small (corresponding to a rare event), $\text{Var}(\hat{\theta})$ is also small, which might be taken to indicate good performance.
- ▶ Not so! For example, if $p = 10^{-6}$ and $n = 10^3$, then $\text{Var}(\hat{\theta})$ is approximately 10^{-9} , i.e. very small.
- ▶ However, with high probability not one of the 1000 binary trials will have resulted in a success, indicating that we are estimating the rare event probability θ by 0!
- ▶ Not what is required.

Rare events, continued

- ▶ A more appropriate measure of performance in rare event simulation is to use *relative error*

$$RE(\hat{\theta}) = \frac{\sqrt{MSE(\hat{\theta})}}{p},$$

where $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ is the mean square error of $\hat{\theta}$.

- ▶ It is worth emphasizing that estimation of rare event probabilities by simulation is a hard problem, essentially because of the problem indicated on the previous slide: unless one does something clever, one may end up estimating the rare event probability by 0.
- ▶ However, there is a growing literature on this topic, due to its obvious importance.
- ▶ Most methods use one of two methods: *importance sampling* and *splitting*.
- ▶ Both involve a process known as “change of measure”.

Importance sampling

- ▶ Simulating X directly from f may be difficult, or it may be feasible but not lead to an adequate procedure.
- ▶ Suppose we can find a distribution g to use as an alternative to f . Then we can estimate θ by

$$\hat{\theta}_g = \frac{1}{n} \sum_{i=1}^N h(X_i) \frac{f(X_i)}{g(X_i)},$$

where now the X_1, \dots, X_N are realisations from g , not f .

- ▶ To see why $\hat{\theta}_g$ is reasonable, note that

$$\theta = \int_{x \in \mathcal{X}} h(x) f(x) dx = \int_{x \in \mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx,$$

which is well-defined provided $f(x) = 0$ whenever $g(x) = 0$.

- ▶ This approach is known as *importance sampling*.
- ▶ This simple idea has a great deal of mileage and is very widely used (not just in rare event simulation).
- ▶ The technique known as *splitting*, used in computational physics, is an alternative “change of measure” technique.

Statistical Hypothesis testing

- ▶ We consider a null hypothesis, H_0 say, versus an alternative hypothesis, H_A say.
- ▶ Hypotheses are expressed in terms of unknown population parameters.
- ▶ For example, if μ is the population mean, we might be interested in testing

$$H_0 : \mu = 3.1 \text{ units} \quad \text{versus} \quad H_A : \mu \neq 3.1 \text{ units.}$$

- ▶ To perform the test, collect data according to a suitable sampling design, and decide between H_0 and H_A on the basis of the test.
- ▶ Well-known examples of statistical tests are the t -test, Wilcoxon test, F -test and χ^2 test.

Errors in hypothesis testing

- ▶ Note that there are two types of “correct” outcome: accepting H_0 when it is true, and rejecting H_0 when it is false.
- ▶ Similarly, two types of error can arise:

Type I error: reject H_0 when it is true

and

Type II error: accepting H_0 when it is false

- ▶ By convention, the type I error is tightly controlled and is regarded as more serious than a type II error.
- ▶ Again by convention, tests are often set up so that the probability of a type I error is 0.05.
- ▶ Until comparatively recently, hypothesis testing was mainly applied in situations in which there was one hypothesis, or a small number of hypotheses, to be tested.

Multiple hypothesis testing and FDR

- ▶ In a number contemporary scientific problems, it is important to be able to test large numbers of hypotheses simultaneously.
- ▶ For example, in microarray data one may wish to examine which of many thousands of genes are “expressed”.
- ▶ At the same time, the number of “cases”, or sample size, may be relatively small, e.g. $n = 50$.
- ▶ In this situation, how can keep a tight control on the number of null hypotheses that we falsely reject? In other words, how do we control the “collective” type I error?
- ▶ For future reference, a p -value is defined as the probability, calculated under the null hypothesis H_0 , of “observing something at least as extreme as what was actually observed”.

Multiple hypothesis testing and FDR

- ▶ FDR is short for *false discovery rate*.
- ▶ To specify this quantity, define

n_{11} = #times the null hypothesis was correctly accepted;

n_{21} = #times the null hypothesis was incorrectly accepted;

n_{12} = #times the null hypothesis was incorrectly rejected;

and

n_{22} = #times the null hypothesis was correctly rejected.

- ▶ Note that $n = n_{11} + n_{21} + n_{12} + n_{22}$ is the total number of hypotheses to be tested, and $n_{12} + n_{22}$ is the total number of hypotheses that were rejected.
- ▶ The FDR is given by

$$FDR = E \left[\frac{n_{12}}{n_{12} + n_{22}} \right].$$

Benjamini and Hochberg approach

- ▶ Benjamini and Hochberg (1995) show that there is a simple way to control the false discovery rate.
- ▶ Let p_j be the p -value corresponding to the j th test ($j = 1, \dots, n$).
- ▶ Let

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$$

denote the ordered p -values.

- ▶ Let k be the largest j for which

$$p_{(j)} \leq \frac{j\alpha}{n\beta},$$

where β is the proportion of cases for which the null hypothesis holds.

- ▶ The Benjamini-Hochberg approach is to reject the null hypothesis in all cases corresponding to $p_{(1)}, \dots, p_{(k)}$, and to accept the null hypothesis in all other cases.
- ▶ The resulting procedure has FDR less than or equal to α .

Benjamini and Hochberg, continued

- ▶ In practice, β , the proportion of cases in which the null hypothesis holds, is usually unknown.
- ▶ In many situations, it is possible to estimate β in a reasonable way.
- ▶ Alternatively, one can replace β by its upper bound of 1. This results in a *conservative* procedure.

Further comments

- ▶ The Benjamini-Hochberg approach has generated a massive amount of interest in recent years.
- ▶ There are now many variants of this approach, including Bayesian and empirical Bayes approaches to multiple hypothesis testing.

Comments and discussion

- ▶ The ideas underlying the simulation techniques described above are simple, but application of these ideas in complex settings typically requires a high level of ingenuity and technical skill.
- ▶ There will always be some trade-off between the sensitivity and specificity of any detection procedure. The use of a formal decision theory framework, involving the use of loss functions, may (or may not) be a useful tool when making this trade-off.
- ▶ Other topics which may (or may not) be relevant include extreme value theory and large deviation theory.