Estimating the probability of a rare event

Laurens de Haan
Erasmus University Rotterdam,
P.O. Box: 1738,
3000 DR Rotterdam,
The Netherlands.

Ashoke Kumar Sinha
Tilburg University,
P.O.Box: 90153,
5000 LE Tilburg,
The Netherlands.

A vast land-pace in the Netherlands is below the average sea level. So to protect the country from flood, dikes are built along the sea coast. Extreme wave-height and still water level are two very important factors for causing flood along the sea coast. Wave-heights and still water levels during high tide have been monitored without any interruption and in a reliable way at several stations along the Dutch sea coast. One such place is *Pettemer zeedijk*. The scientists concerned with the safety of Pettemer zeedijk found out that a calamity can occur if the wave-height and still water level (both measured in meter) satisfy the following relation: $0.3 \times$ wave-height [m.] + sea-level [m.] ≥ 7.6 [m.].

The region $C=\{(s,t):0.3s+t \geq 7.6\}$, where s = wave-height and t = sea-level, is a possible failure region. Wave-height and sea level were recorded at Pettemer zeedijk during 828 independent storm events. Fortunately, up to now we did not observe any wave-height and sea-level combination which falls into the region C. Our main problem is to estimate the probability that a future storm can cause a wave-height and sea-level combination which falls in the failure region C.

We can formulate the problem mathematically as follows: Let $\{(X_i,Y_i);i=1,\ldots,n\}$ be a sample from the bivariate distribution function F. On the basis of the sample, we want to estimate $p:=Pr((X,Y)\in C)$, the failure probability.

Since so far we did not have any observation, which falls into the failure region C, we cannot use the empirical distribution function to estimate p. Since none of the observations fall into C, in first approximation the probability p must be less than $1/n$. Hence to estimate p, one has to extrapolate outside the range of the sample. In a finite-sample context this problem cannot be solved without assuming a certain parametric model (cf. for example Smith, Tawn and Yuen (1990), Coles and Tawn (1991), Joe, Smith and Weissman (1992), Coles and Tawn (1994)). But a fixed model, which may fit very well to the sample, may not reflect the behaviour of the variables beyond the range of the observation. So it may turn out to be unsuitable for extrapolation. Hence we do not consider a finite-sample framework and rather set up an asymptotic framework where the sample-size goes to infinity. For the purpose of extrapolation, we need to assume certain smoothness conditions of F near the endpoints of both variables. Thus we shall assume some extra condition on the distribution function F in the field of extreme-value theory.

Now the fact that none of the observations is close to the failure region is an essential feature of the problem which we want to retain when applying asymptotic theory as we will do. So

the inequality $n \times Pr((X,Y) \in C) < 1$ forces us to assume that in fact, when applying asymptotic theory, the set C depends on n (notation: C_n) and that the sequence $p_n := Pr((X,Y) \in C_n)$ tends to zero as $n \to \infty$. In this article, we propose an estimate of p_n in an asymptotic set-up. We shall also construct a confidence interval of the failure probability.

REFERENCES

Coles, S.G. and Tawn, J.A. (1991). Modelling extreme multivariate events. J. Roy. Statist. Soc. series B.

Coles, S.G. and Tawn, J.A. (1994). Statistical methods for multivariate extremes: an application to structural design. Appl. Statist. 43, 1-48.

Joe, H., Smith, R.L. and Weissman, I. (1992). Bivariate threshold methods for extremes. J. Roy. Statist. Soc. series B 45, 171-183.

Smith, R.L., Tawn, J.A. and Yuen, H.K. (1990). Statistics of multivariate extremes. Int. Statist. Rev 58, 47-58.