Exercises

N.3 **Biggest of bunch: Gumbel.**[1] (Mathematics, Statistics, Engineering) ③

Much of statistical mechanics focuses on the average behavior in an ensemble, or the mean square fluctuations about that average. In many cases, however, we are far more interested in the extremes of a distribution.

Engineers planning dike systems are interested in the highest flood level likely in the next hundred years. Let the high water mark in year j be H_j. Ignoring long-term weather changes (like global warming) and year-to-year correlations, let us assume that each H_j is an independent and identically distributed (IID) random variable with probability density $\rho_1(H_j)$. The *cumulative distribution function* (cdf) is the probability that a random variable is less than a given threshold. Let the cdf for a single year be $F_1(H) = P(H' < H) = \int^H \rho_1(H')\,\mathrm{d}H'$.

(a) *Write the probability $F_N(H)$ that the highest flood level (largest of the high-water marks) in the next $N = 1000$ years will be less than H, in terms of the probability $F_1(H)$ that the high-water mark in a single year is less than H.*

The distribution of the largest or smallest of N random numbers is described by *extreme value statistics* [10]. Extreme value statistics is a valuable tool in engineering (reliability, disaster preparation), in the insurance business, and recently in bioinformatics (where it is used to determine whether the best alignments of an unknown gene to known genes in other organisms are significantly better than that one would generate randomly).

(b) *Suppose that $\rho_1(H) = \exp(-H/H_0)/H_0$ decays as a simple exponential ($H > 0$). Using the formula*

$$(1 - A) \approx \exp(-A) \quad \text{small A} \tag{N.1}$$

show that the cumulative distribution function F_N for the highest flood after N years is

$$F_N(H) \approx \exp\left[-\exp\left(\frac{\mu - H}{\beta}\right)\right]. \tag{N.2}$$

for large H. (Why is the probability $F_N(H)$ small when H is not large, at large N?) What are μ and β for this case?

The constants β and μ just shift the scale and zero of the ruler used to measure the variable of interest. Thus, using a suitable ruler, the largest of many events is given by a Gumbel distribution

$$\begin{aligned} F(x) &= \exp(-\exp(-x)) \\ \rho(x) &= \partial F/\partial x = \exp(-(x + \exp(-x))). \end{aligned} \tag{1}$$

How much does the probability distribution for the largest of N IID random variables depend on the probability density of the individual random variables? Surprisingly little! It turns out that the largest of N Gaussian random variables also has the same Gumbel form that we found for exponentials. Indeed, any probability distribution that has unbounded possible values for the variable, but that decays faster than any power law, will have extreme value statistics governed by the Gumbel distribution [5, section 8.3]. In particular, suppose

$$F_1(H) \approx 1 - A\exp(-BH^\delta) \tag{N.4}$$

as $H \to \infty$ for some positive constants A, B, and δ. It is in the region near $H^*[N]$, defined by $F_1(H^*[N]) = 1 - 1/N$, that F_N varies in an interesting range (because of eqn N.1).

(c) *Show that the extreme value statistics $F_N(H)$ for this distribution is of the Gumbel form (eqn N.2) with $\mu = H^*[N]$ and $\beta = 1/(B\delta\, H^*[N]^{\delta-1})$.* (Hint: Taylor expand $F_1(H)$ at H^* to first order.)

The Gumbel distribution is *universal*. It describes the extreme values for any unbounded distribution whose tails decay faster than a power law.[2] (This is quite analogous to the central limit theorem, which

[1] New exercise supplementing *Statistical Mechanics: Entropy, Order Parameters, and Complexity* by James P. Sethna, copyright Oxford University Press, 2007, page 3. A pdf of the text is available at pages.physics.cornell.edu/sethna/StatMech/ (select the picture of the text). Hyperlinks from this exercise into the text will work if the latter PDF is downloaded into the same directory/folder as this PDF.

[2] The Gumbel distribution can also describe extreme values for a bounded distribution, if the probability density at the boundary goes to zero faster than a power law [10, section 8.2].

shows that the normal or Gaussian distribution is the universal form for sums of large numbers of IID random variables, so long as the individual random variables have non-infinite variance.)

The Gaussian or standard normal distribution $\rho_1(H) = (1/\sqrt{2\pi})\exp(-H^2/2)$, for example, has a cumulative distribution $F_1(H) = (1/2)(1 + \mathrm{erf}(H/\sqrt{2}))$ which at large H has asymptotic form $F_1(H) \sim 1 - (1/\sqrt{2\pi}H)\exp(-H^2/2)$. This is of the general form of eqn N.4 with $B = 1/2$ and $\delta = 2$, except that A is a slowly varying function of H. This slow variation does not change the asymptotics. Hints for the numerics are available in the computer exercises section of the text Web site [8].

(d) *Generate* $M = 10000$ *lists of* $N = 1000$ *random numbers distributed with this Gaussian probability distribution. Plot a normalized histogram of the largest entries in each list. Plot also the predicted form* $\rho_N(H) = \mathrm{d}F_N/\mathrm{d}H$ *from part (c).* (Hint: $H^*(N) \approx 3.09023$ for $N = 1000$; check this if it is convenient.)

Other types of distributions can have extreme value statistics in different universality classes (see Exercise N.8). Distributions with power-law tails (like the distributions of earthquakes and avalanches described in Chapter 12) have extreme value statistics described by *Fréchet distributions*. Distributions that have a strict upper or lower bound[3] have extreme value distributions that are described by Weibull statistics (see Exercise N.4).

[3] More specifically, bounded distributions that have power-law asymptotics have Weibull statistics; see note 2 and Exercise N.4, part (d).