AAPT PHYSICS EDUCATION | AMERICAN JOURNAL of PHYSICS

"Weather" records: Musings on cold days after a long hot Indian summer

B. Schmittmann and R. K. P. Zia

Citation: American Journal of Physics **67**, 1269 (1999); doi: 10.1119/1.19114
View online: http://dx.doi.org/10.1119/1.19114
View Table of Contents: http://scitation.aip.org/content/aapt/journal/ajp/67/12?ver=pdfcov
Published by the American Association of Physics Teachers

''Weather'' records: Musings on cold days after a long hot Indian summer

B. Schmittmann and R. K. P. Zia
Center for Stochastic Processes in Science and Engineering and Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0435

(Received 27 April 1999; accepted 24 May 1999)

We present a simple introduction to the statistics of extreme values. Motivated by a string of record high temperatures in December 1998, we consider the distribution, averages, and lifetimes for a simplified model of such ''records.'' Our data are sequences of independent random numbers all of which are generated from the same probability distribution. A remarkable universality emerges: A number of results, including the lifetime histogram, are universal, that is, independent of the underlying distribution.

I. INTRODUCTION AND MOTIVATION

In December 1998, in the aftermath of El Niño and its companion, La Niña, the weather in the Roanoke, Virginia, area was unusually mild. Weather data have been collected in Virginia since 1934, and record highs and lows for any particular day are known. As part of the daily weather forecasts, local TV stations report these record values and compare them to the highest and lowest temperature values of the day. Remarkably, during the nine days from November 29 to December 7, 1998, the previous record highs were broken five times and tied once![1] One might wonder, as we did, how frequently such a series of records could occur. When only few weather data are available, such as in the early years of record keeping, it is obviously quite easy to experience new extremes. However, as the data sets become larger, the record highs (and lows) are pushed to higher (lower) values, so that the setting of a new record becomes a less frequent event. Thus, we began exploring questions such as: How probable was it to set a new record in 1998, 64 years after the records began? How do record highs increase with time? How long do records typically survive before they are broken?

Not surprisingly, similar questions have been posed before. It appears that the earliest studies are due to N. Bernoulli, who analyzed life expectancies in 1709.[2] Later, flood control and structural safety issues were considered, to name just a few of the numerous applications. Beginning in the 1920s, mathematical techniques were developed and the study of records, or extremes, became known as ''extreme value statistics,''[2] an active area of research. In this paper, we provide a pedagogical introduction to some of the basic results. We begin in Sec. II with the simplest model for records and a concise statement of the problem. There is no attempt to address real records. As a result of complex physical processes, the statistics of real weather records is undoubtedly far more complex than that generated by simple random numbers. The quotation marks in the title should remind the reader that actual weather records are *not* the subject of this article. Section III is devoted to the full distribution function for the model records, their averages, and standard deviations. Next, we discuss the record lifetimes and derive the associated distribution in Sec. IV. Although we focus on ''record highs,'' a completely analogous line of reasoning can be pursued for ''record lows.''[2] In the final section, we turn to more general questions and conclude by listing a number of problems which, to the best of our knowledge, are still unsolved. Some technical details are provided in the Appendix.

II. A SIMPLE MODEL FOR RECORDS

Our much simplified model for the physical data (temperatures, water levels, etc.) is based on a probability density $p(x)$ for a real variable x. For example, if we are considering temperatures, $p(x)dx$ might be the probability that the temperature at noon, on a specific day of the year, takes a value between x and $x+dx$. The ''data'' in our simplified model are just a sequence of random numbers:

$$\{x(1),x(2),x(3),\ldots,x(t),\ldots\}. \tag{1}$$

In keeping with the language of weather records, we refer to the (integer) label of these numbers as the ''time'' (or ''year''), $t=1,2,3,\ldots$. Each of the x's is drawn from the *same distribution* $p(x)$. Thus, our data form a set of independent, identically distributed random variables. This condition is probably the most serious shortcoming when applied to physical reality, where major correlations or variations can be expected. For example, it precludes ''global warming,'' a situation in which the underlying distribution $p(x)$ is a function of time.

Before discussing the sequence and the records, we discuss some details about the distribution $p(x)$. It may be defined over the entire real axis or restricted to an interval, that is, it may have infinite or finite support. For simplicity, we require it to be *continuous and positive*, which excludes, for example, dice throwing. As a result, we can ignore the possibility of ties. Of course, it must be normalized, that is, $\int p(x)dx=1$. Here and in the following, any integral limits that are not explicitly specified are to be taken as the appropriate (upper and/or lower) bounds of the support. Because the distribution $p(x)$ reflects how we model our data set, we refer to it as the ''model distribution,'' or simply ''the model.'' In particular, we will investigate to what extent our results for record distributions and lifetimes depend on the underlying model. For later reference, we also introduce another distribution related to $p(x)$ by

$$q(R)\equiv\int^{R} p(x)dx. \tag{2}$$

cumulativa

Clearly, $q(R)$ is the probability that a random number, drawn from the distribution $p(x)$, will not exceed R. Thus, $q(R)$ is a monotonic function, varying from 0 to 1. The reader familiar with the generation of random numbers on a computer will recognize that the inverse of this function, $R(q)$, is the inverse transform method for generating random numbers for an arbitrary density $p(x)$ from the uniform distribution on the interval $[0,1]$.[3] Following Galambos,[2] we

call $q(R)$ the ''common distribution function.''

Returning to the sequence (1), we define the ''record'' $R(t)$ as the *largest* number in a string of length t:

$$R(t)\equiv\max\{-\infty,x(1),x(2),x(3),\ldots,x(t)\}$$
$$\equiv\sup\{x(1),x(2),x(3),\ldots,x(t)\}, \tag{3}$$

where we have included $R(0)=-\infty$ as the (arbitrary) initial value. After generating the next new number $x(t+1)$, we determine whether the record has been broken, according to

$$R(t+1)=\begin{cases} R(t) & \text{if } x(t+1)\leqslant R(t) \\ x(t+1) & \text{if } x(t+1)>R(t). \end{cases} \tag{4}$$

This procedure is continued until we have obtained a sequence of length N. Of course, the records $R(t)$ themselves are stochastic quantities. So, we can define a probability density for R, $P(R,t)$, so that the probability for finding the record to lie between R and $R+dR$ is $P(R,t)dR$. Our goal is to determine, given the underlying model distribution $p(x)$, $P(R,t)$ or more simply, the average record:

$$\langle R(t)\rangle=\int dRRP(R,t). \tag{5}$$

We first consider computer simulations of these records. To study their statistics, we generate a large number M of sequences: $R_i(t), i=1,\ldots,M$. Based on this ensemble of sequences, we can define the average record as a function of time, $\langle R(t)\rangle=\Sigma_{i=1}^{M}R_i(t)/M$. For example, consider $M=2$ sequences of $N=10$ random numbers (given to three digits for simplicity): $x_1(t)=\{0.686, 0.314, 0.456, 0.808, 0.002, 0.972, 0.937, 0.141, 0.706, 0.970\}$ and $x_2(t)=\{0.104, 0.900, 0.713, 0.464, 0.260, 0.986, 0.259, 0.058, 0.778, 0.886\}$. The corresponding sequences of records are

$$R_1(t)=\{0.686, 0.686, 0.686, 0.808, 0.808, 0.972, 0.972, 0.972, 0.972, 0.972\}, \tag{6}$$

$$R_2(t)=\{0.104, 0.900, 0.900, 0.900, 0.900, 0.986, 0.986, 0.986, 0.986, 0.986\}. \tag{7}$$

Hence, $\langle R(t)\rangle=\{0.395, 0.793, 0.793, 0.854, 0.854, 0.979, 0.979, 0.979, 0.979, 0.979\}$. Because the random numbers are uniformly distributed between 0 and 1, it is not surprising to see that the average record bumps up against one. To get a good grasp of this process, we suggest that the reader should attempt $M=100$ sequences of $N=300$.

For the computer simulations presented in this article, we used a high quality random generator (RAN2 from *Numerical Recipes*[4]) and averaged over $M=10^5$ sequences of length $N=10^3$. This procedure requires just a few minutes running on a 450-MHz Pentium II running Linux, and produces excellent statistics. In fact, with 10^5 sequences, it is possible to generate a reasonable picture of the entire probability distribution for the records. A theoretical approach to the averages will be our next task.

III. THE RECORD PROBABILITY DISTRIBUTION

A. The evolution equation and its solution

Our goal in this section is to find an analytical form for the probability density $P(R,t)$. We will do this recursively, by assuming that we know the value of the record, say R', at time $t-1$. Then, we seek the *conditional* probability, $P(R,t|R',t-1)$, that the record, at time t, has the value R, provided it had the value R' at time $t-1$. Fortunately, it is very easy to write down this quantity. Clearly, it vanishes for $R<R'$, because the new record cannot be smaller than the old one. This consideration leaves us with two possibilities: Either the old record stays the same so that $R=R'$, or it is broken, resulting in the new value $R>R'$. The former is the case if the new random number, $x(t)$, does not exceed R'. From Eq. (2), we see that this case occurs with probability $q(R)$. In contrast, if $x(t)$ exceeds R', then it sets the new record R. This latter case occurs with probability $p(R)$. Hence, the conditional probability is

$$P(R,t|R',t-1)=\begin{cases} 0, & R<R' \\ q(R) \text{ for } & R=R' \\ p(R), & R>R' \end{cases}$$
$$=q(R)\delta(R-R')+p(R)\Theta(R-R'). \tag{8}$$

The Heavyside step function Θ is unity if its argument is positive and zero otherwise. Its derivative is just the Dirac delta function δ, so that the two terms in Eq. (8) can be combined into

$$P(R,t|R',t-1)=\frac{\partial}{\partial R}[q(R)\Theta(R-R')]. \quad \text{ok} \tag{9}$$

Because we have exhausted all possibilities for R, the conditional probability must be normalized with respect to an integration over R. This condition is easily checked: Because P is a total derivative, its integral is just $q\Theta$ evaluated at the limits. So, we have $\int dRP(R,t|R',t-1)=1$.

From the *conditional* probability, it is a simple step to arrive at our main target: the record probability density (the ''record'' distribution) $P(R,t)$: Chapman Kolmogorov equation

$$P(R,t)=\int dR'P(R,t|R',t-1)P(R',t-1). \tag{10}$$

Substituting the explicit form (9) for the conditional probability, we obtain a recursion relation for P:

$$P(R,t)=\int dR'\frac{\partial}{\partial R}[q(R)\Theta(R-R')]P(R',t-1)$$
$$=\frac{\partial}{\partial R}\left[q(R)\int dR'\Theta(R-R')\right]P(R',t-1)$$
$$=\frac{\partial}{\partial R}q(R)\int^{R}dR'P(R',t-1). \tag{11}$$

This form is still slightly unwieldy due to the integration. Let us define the ''barrier'' distribution $Q(R,t)$ to be the probability that at time t the record is at R or lower:

$$Q(R,t)\equiv\int^{R}dR'P(R',t), \tag{12}$$

so that

$$P(R,t)=\frac{\partial}{\partial R}Q(R,t). \tag{13}$$

From Eqs. (11) to (13), the function Q satisfies a very simple recursion equation, namely,

$$Q(R,t)=q(R)Q(R,t-1). \tag{14}$$

Equation (14) has a simple interpretation. Recall that $q(R)$ is the probability that the next random number is less than or equal to R, and $Q(R,t)$ is the probability that at time t, the record has *not exceeded* R. The product of the two gives the probability that the record remains unbroken after the next time step.

The recursion relation (14) is easily solved:

$$Q(R,t)=[q(R)]^tQ(R,0)=[q(R)]^t, \tag{15}$$

because $Q(R,0)=1$ for all $R>-\infty$. We can deduce two important properties from the general solution (15) for the barrier distribution Q. Because $q(R)$ is a monotonic function, so is $Q(R)$. This behavior of $Q(R,t)$ implies that at any given time t, it is more difficult to go over a higher barrier (break a record). On the other hand, because $q<1$ for any fixed $R<1$, $Q(R,t)$ *decreases* with t, which implies that any given record can be broken, provided we wait long enough. We emphasize that these ''reasonable'' results are completely *independent* of the details of the underlying distribution $p(x)$.

From Eq. (15), the record distribution follows:

$$P(R,t)=\frac{\partial}{\partial R}Q(R,t)=t[q(R)]^{t-1}\frac{\partial q(R)}{\partial R}$$
$$=t[q(R)]^{t-1}p(R). \tag{16}$$

Once $P(R,t)$ is known, we can compute average values for the records, $\langle R(t)\rangle$, using its definition (5), as well as the standard deviations and all higher moments. For later reference, we provide a simplified representation of the integral in Eq. (5). Substituting Eq. (16) into Eq. (5), we have $\langle R(t)\rangle = \int dRRt[q(R)]^{t-1}p(R)$. Next, recall that $q(R)$ is monotonic, so that it can be inverted uniquely to give the function $R(q)$. Then change the integration variable from R to q, using $p(R)dR=dq$. The result is

$$\langle R(t)\rangle=t\int_0^1 dqR(q)q^{t-1}. \tag{17}$$

Note that the limits of integration are now *unique*, that is, independent of the underlying model.

In Sec. II B, we will illustrate the characteristics of $P(R,t)$ and $\langle R(t)\rangle$ with two simple, explicitly solvable forms of the distribution p.

B. Two exactly solvable examples: Flat and purely exponential distributions

We illustrate our results by considering two particularly simple cases: a flat distribution

$$p(x)=1 \quad \text{for} \quad 0\leqslant x\leqslant 1 \tag{18}$$

and a pure exponential

$$p(x)=e^{-x} \quad \text{for} \quad 0\leqslant x. \tag{19}$$

These distributions differ significantly in that the former has an *upper bound* for the allowed values of the data. As a result, the possible record values are also bounded. In contrast, the latter distribution extends to infinity, setting no limits on the possible records. These two simple distributions will yield rather different, but hopefully generic behavior for bounded versus unbounded models.

The flat distribution (18) corresponds to data whose values are equally probable over a given interval. From Eq. (15), we find $q(R)=R$ (for $0\leqslant R\leqslant 1$) and

$$Q(R,t)=[q(R)]^t=R^t, \tag{20}$$

so that

$$P(R,t)=\partial Q/\partial R=tR^{t-1}. \tag{21}$$

Both results are easily interpreted. The barrier distribution $Q(R,t)$ displays explicitly the general properties discussed above: increasing with R at fixed t and decreasing with t at fixed $R<1$. In contrast, the record distribution $P(R,t)$ displays a *maximum in* t for a fixed R. For early times, the probability to find the record at R is low because this value has not yet been reached, whereas, for late times, it has already been exceeded.

We can also study $P(R,t)$ as a function of R for a fixed t. The maximum value always occurs at $R=1$, increasing linearly with time. If we apply the normalization condition $\int PdR=1$, the width of the peak must narrow with time. In other words, the late time records are ''piled up'' just below the highest allowed value (unity in this case). Let us investigate how this feature is reflected in the *average* record. Using Eq. (5), we find easily that

$$\langle R(t)\rangle=t\int_0^1 dRRR^{t-1}=\frac{t}{t+1}. \tag{22}$$

As expected, $\langle R(t)\rangle$ increases monotonically as a function of time, reaching its upper bound at $t=\infty$:

$$\lim_{t\to\infty}\langle R(t)\rangle=1. \tag{23}$$

Of course, its *rate* of increase must vanish in this limit. In this case, the asymptotic rate is t^{-2}. This behavior is illustrated in Fig. 1(a), which shows that the exact result, Eq. (22), is in excellent agreement with Monte Carlo data averaged over 10^5 sequences with 1000 entries each.

Several of these properties are generic in the following sense. If the underlying distribution p has an upper bound, that is, $p(x)=0$ for $x\geqslant B$, then $\lim_{t\to\infty}\langle R(t)\rangle=B$. Not surprisingly, the behavior of p near B will dictate the asymptotics. For example, if we assume $p(x)\to\mu k(B-x)^{\mu-1}$, so that $q(R)\to 1-k(B-R)^{\mu}$ for x, then $R\to B$, and we can show that the difference $B-\langle R(t)\rangle\to\Gamma(1/\mu)(kt)^{-1/\mu}$, which is a generalization of the flat case.

We next turn to the purely exponential distribution (19) and its associated distribution $q(R)=1-p(R)$. The two characteristic distribution functions are

$$Q(R,t)=[1-e^{-R}]^t \tag{24}$$

and

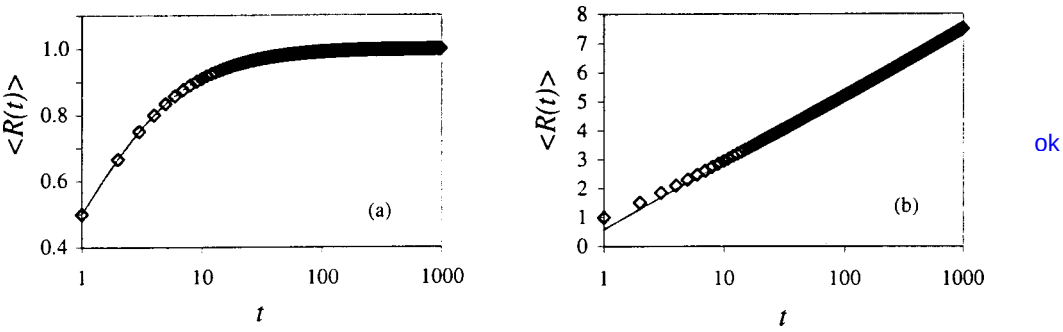
$$P(R,t)=te^{-R}[1-e^{-R}]^{t-1}. \tag{25}$$

Many properties are qualitatively similar to the flat case. One difference is that the position, $R_0(t)$, of the maximum in $P(R,t)$ increases with time indefinitely:

$$R_0(t)=\ln t. \tag{26}$$

Once again, this behavior is reflected in the average record. Using Eq. (17) and deferring the details to Appendix A, we obtain

$$\langle R(t)\rangle=\sum_{k=1}^{t}\frac{1}{k}, \tag{27}$$



ok

Fig. 1. Average record $\langle R(t)\rangle$ vs time t, for (a) flat and (b) exponential distributions. The diamonds are Monte Carlo simulation data; the solid lines are the theoretical results, Eqs. (22) and (29), respectively.

which is obviously a monotonically increasing function of time. To exhibit the asymptotic behavior for large times, we can write Eq. (27) in more compact notation:

$$\langle R(t)\rangle = \psi(t+1) - \psi(1), \tag{28}$$

where ψ is the digamma function with known asymptotic behavior.[5] As a result,

$$\langle R(t)\rangle \simeq \ln t + \gamma + O(1/t) \quad \text{for} \quad t\to\infty, \tag{29}$$

where $\gamma \simeq 0.5772$. Note that the *rate* of increase also vanishes with time, but with a slower decay, t^{-1}. In Fig. 1(b), we show the excellent agreement of the asymptotic result (29) with the Monte Carlo data.

As before, we can consider the behavior of the average record for more general unbounded distributions $p(x)$. Here, the argument rests on a saddle-point approximation for $P(R,t)$ around its maximum value, $R_0(t)$. Asymptotically, the average can be approximated by the position of the maximum. Because $R_0(t)$ becomes very large for late times, the asymptotics should be controlled by the behavior of $p(x)$ for large x. Some examples of possible asymptotic behaviors of unbounded ps and their associated saddle points $R_0(t)$ are

(1) a power law $p\sim 1/x^{\alpha}$ (with $\alpha>2$ to ensure that $\langle R\rangle$ exists), resulting in $R_0(t)\sim t^{1/(\alpha-1)}$;
(2) an exponential $p\sim e^{-\mu x}$, giving $R_0(t)\sim \ln t$;
(3) a Gaussian $p\sim e^{-\mu R^2}$ for which $R_0(t)\sim \sqrt{\ln t}$.

We invite the reader to do simulations for, say, $p(x)=2/x^3$, $x\in[1,\infty]$ and compare their results to the predictions in item (1). We would also like to caution the reader that the approach to asymptopia for the Gaussian is *extremely* slow. In particular, we find that this regime lies beyond $t=10^6$.

IV. THE DISTRIBUTION OF RECORD LIFETIMES

Next, we turn to a key question in the study of floods or earthquakes. What is the typical time span between two large events? At the practical level, how much time do we have to construct dams or to repair levees before the next record-breaking flood? Of course, our study will not provide an exact time span until the next disaster; it will only give an estimate of how long a record might survive. We refer to the time span between the setting of a record and its subsequent breaking as its ''lifetime'' (or ''record time''[6]). In the following, we investigate the distribution of these record lifetimes. Specifically, we will consider data sequences with N entries. For each sequence, we will identify the associated records and their lifetimes. By analyzing a large number of data sequences, we can compile a histogram, $T_N(m)$, of how likely a record would survive for a time span m.

A. A tree of lifetimes

Let us introduce a tree-like structure to represent all possible histories of records. We begin by generating a data sequence $\{x(1),x(2),x(3),\ldots,x(N)\}$. Because we only need to know *where* the records occur in this sequence, we can associate this sequence with the following binary string. If $x(i)$ is a new record, replace it by the letter R (''record''); otherwise, replace it by L (''lower''). Note that if we had used a discrete underlying distribution p, we would have had to consider the complication of ties. As an example, the sequence $\{0.2,0.4,0.3,0.1,0.6,0.7,0.2,0.4,0.8\}$ is replaced by $\{R,R,L,L,R,R,L,L,R\}$. By convention, the first entry is always R. If a record is established at time t and broken at time $t+m$, then that record's lifetime is defined to be m. Clearly, a binary string is much simpler than the original data sequence, but it contains enough information about record lifetimes for us to predict the distribution $T_N(m)$.

The binary strings are easily visualized via a tree structure. Starting from a single vertex (the ''ancestor'') R on the first line ($t=1$), time runs downwards. At time $t=2$ (the second line), our string has two possible entries: R or L. To represent these, we draw two branches from the first line to the second: one to the right (labeled R), and one toward the left (labeled L). Each of these branches ends in a new vertex. These branch again, giving us four vertices in total on the third ($t=3$) line. Continuing this procedure to the Nth level, we find 2^{N-1} vertices. As an illustration, the $N=4$ tree is shown in Fig. 2(a). Clearly, all possible binary strings with four elements appear in this tree, each associated with exactly one vertex on the $t=4$ line.

The record lifetimes associated with a given string are now easily identified: following an L branch means that the current record survives, while choosing the R branch implies that a new record is set. Thus, each vertex can be labeled with the set of record lifetimes $\{\tau_1,\tau_2,\ldots,\tau_k\}$ leading to it. Figure 2(b) shows the ''tree of lifetimes'' for the $N=4$ case shown in Fig. 2(a). Note that the number of entries in these sets varies for different strings. For example, the far right string in Fig. 2(a), where every record is broken at the next time step, gives rise to $\{1,1,1,1\}$, while the far left string corresponds to $\{4\}$: the record is set at $t=1$ and survives the

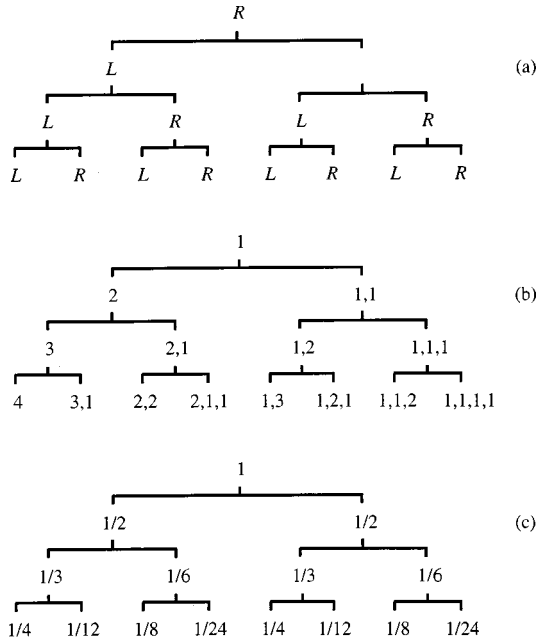


Fig. 2. Tree of records. At the Nth level of the tree are the 2^N possible histories. (a) The binary strings representing the histories of records. (b) The lifetime associated with each string. (c) Probabilities associated with each history, or ''string probabilities.''

next three time steps. A simple recursion relation emerges. From a particular entry at time t to the two ''daughters'' at time $t+1$, the set $\{\tau_1,\tau_2,\ldots,\tau_{k-1},\tau_k\}$ branches into two: $\{\tau_1,\tau_2,\ldots,\tau_{k-1},\tau_k+1\}$ and $\{\tau_1,\tau_2,\ldots,\tau_k,1\}$. Moreover, because any distribution $p(x)$ will generate the same set of binary strings, it is obvious that the associated lifetimes $\{\tau_1,\tau_2,\ldots,\tau_k\}$ are completely *independent* of the underlying model!

B. The string probabilities

To complete the construction of the histogram, we need to find the *string probability*, that is, the probability that a specific string will appear. At the Nth level, each string is associated with a specific vertex, which is uniquely labeled by the set $\{\tau_1,\tau_2,\ldots,\tau_{k-1},\tau_k\}$. Let us denote this probability by $P_N(\tau_1,\tau_2,\ldots,\tau_k)$. At the top level $(t=1)$, this probability is trivial: $P_1(1)=1$. For example, the year record keeping starts, any temperature will be a ''record.'' In the second ''year'' $(t=2)$, there are two vertices: $\{\tau_1=2\}$ and $\{\tau_1=1,\tau_2=1\}$, corresponding to the first record surviving or being broken, respectively. In terms of the original data sequence $\{x(1),x(2)\}$, these possibilities are given by $x(1)>x(2)$ or $x(1)<x(2)$. Let us focus first on the former case for which the probability, $P_2(2)$, is $\int dx_1\int dx_2 p(x_1)p(x_2)\times\Theta(x_1-x_2)$. At first sight, this expression seems to depend on the model distribution $p(x)$. However, recalling Eq. (17), we can transform the integration variables to

$$P_2(2)=\int_0^1 dq_1\int_0^1 dq_2\,\Theta(q_1-q_2)=\int_0^1 dq_1\int_0^{q_1} dq_2=\frac{1}{2}, \tag{30}$$

where we have exploited the monotonicity of $q(x)$ to replace $\Theta(x_1-x_2)$ by $\Theta(q_1-q_2)$. Not only is this integral trivial to compute, it is also *manifestly independent* of the underlying model. In a similar manner, we can compute explicitly that the probability of the latter case, $P_2(1,1)$, is also $\frac{1}{2}$.

Let us illustrate this process for the next year $(t=3)$. For the four possible histories, different combinations of Θ functions appear. Their associated probabilities are

$$\begin{aligned}
P_3(3)&=\int_0^1 dq_1\int_0^1 dq_2\int_0^1 dq_3\,\Theta(q_1-q_2)\Theta(q_1-q_3)=\frac{1}{3},\\
P_3(2,1)&=\int_0^1 dq_1\int_0^1 dq_2\int_0^1 dq_3\,\Theta(q_1-q_2)\Theta(q_3-q_1)=\frac{1}{6},\\
P_3(1,2)&=\int_0^1 dq_1\int_0^1 dq_2\int_0^1 dq_3\,\Theta(q_2-q_1)\Theta(q_2-q_3)=\frac{1}{3},\\
P_3(1,1,1)&=\int_0^1 dq_1\int_0^1 dq_2\int_0^1 dq_3\,\Theta(q_2-q_1)\Theta(q_3-q_2)=\frac{1}{6}.
\end{aligned} \tag{31}$$

In Fig. 2(c), we show all the probabilities in the $N=4$ tree. Note that the sum of all the P_Ns at each level is unity; the probability of having *any* history after N years must be 1.

In Appendix B, we show that the general result for an arbitrary string of any length N is

$$P_N(\tau_1,\tau_2,\ldots,\tau_k)=\frac{1}{\tau_1(\tau_1+\tau_2)(\tau_1+\tau_2+\tau_3)\ldots(\Sigma_1^k\tau_i)}. \tag{32}$$

Note that the last factor is just N.

Another somewhat counterintuitive result concerns $S_N(m)$, the probability for the *last record to survive m* steps (regardless of what happened earlier). Note that the ''last record'' is also the ''highest record,'' because the last record is necessarily larger than all previous records. To say that this record survives m steps means that the values in the rest of the string ($m-1$ of them) are lower. Therefore, in a string of N steps, the last record must have occurred at the $l\equiv(N-m+1)$th step. Thus, to find $S_N(m)$, we ask: What is the probability for the highest record to show up at the lth step? We might guess that the highest record is unlikely to occur near the beginning of the string, because there are many chances for it to be broken later. On the other hand, if we use the language of athletics, we might conclude that the record facing the last athlete may be quite high (given that ''many guys have gone before'') and breaking the record may not be easy. So, perhaps $S_N(m)$ should be peaked in the middle? The surprise is that $S_N(m)$ is *independent of m* (or equivalently, l)! In other words, the highest record *may occur at any step with equal probability*. The proof is in Appendix C.

C. The universal lifetime distribution

Finally, we turn to the (unnormalized) lifetime distribution, $T_N(m)$. We illustrate the $N=4$ case by explicit calculations and relegate the general discussion to Appendix D. Combining the string probabilities with the lifetimes, we obtain

$$\begin{aligned}
T_4(4)&=P_4(4)=\tfrac{1}{4},\\
T_4(3)&=P_4(3,1)+P_4(1,3)=\tfrac{1}{12}+\tfrac{1}{4}=\tfrac{1}{3},\\
T_4(2)&=2P_4(2,2)+P_4(2,1,1)+P_4(1,2,1)+P_4(1,1,2)\\
&=\tfrac{1}{2},
\end{aligned} \tag{33}$$

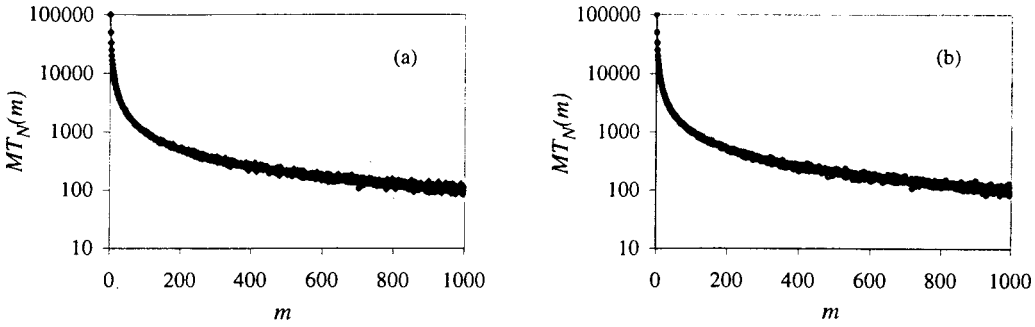


Fig. 3. Histogram of record lifetimes, for (a) flat and (b) exponential distributions. The normalization factor $M=10^5$ is the number of sequences that have been averaged. The theoretical line is barely visible behind the data points.

$$T_4(1)=P_4(3,1)+2P_4(2,1,1)+P_4(1,3)+2P_4(1,2,1)$$
$$+2P_4(1,1,2)+4P_4(1,1,1,1)=1.$$

Focusing on a particular binary string, we recall that the associated record lifetimes $\{\tau_1,\tau_2,\ldots,\tau_k\}$ as well as the string probability are model independent. As a result, the lifetime distribution is *universal*, that is, its form does not depend on the underlying model distribution $p(x)$. Moreover, the result for $T_4(m)$ suggests a surprisingly simple form for the lifetime histogram, namely,

$$T_N(m)=1/m. \tag{34}$$

The general proof can be found in Appendix D. Remarkably, T_N is not only universal, but also is independent of N, that is, the length of the original data sets. So, the probability for a record to survive, for example, five time steps is the same, no matter how much data we accumulate. Phrased differently, this result is not surprising at all: Because $T_N(m)$ is also, roughly speaking, the probability for the next huge disaster to strike after m time steps, it would be very disturbing if that probability depended on the length of our data sets: that would imply that we could avoid or court disaster by just continuing to take data. Clearly, such a result would be nonsensical. Nobody would believe that, by observing the weather, we can change it!

Although the lack of an N dependence can be understood in this way, the universality with respect to the underlying model $p(x)$ is a much stronger statement: As long as *all* underlying data arise from the *same* distribution, the functional form of this distribution is irrelevant. Flat, exponential or Gaussian p's all generate identical lifetime histograms. In Fig. 3, we show Monte Carlo data for the lifetime histograms of flat and exponential $p(x)$. They are indistinguishable, apart from statistical fluctuations. The agreement with the theoretically expected $T_N(m)=1/m$ is again excellent.

V. CONCLUDING REMARKS

We have presented a simple introduction to the statistics of extremes. Although we motivated the discussion by reference to weather records, the model we studied is much simplified. Generating a sequence of N random numbers, all selected from the same underlying distribution $p(x)$, we keep track of the ones which are higher than all the preceding numbers—''record highs.'' Studying the statistics of such strings, we identify several universal features, that is, properties which are independent of the underlying distribution p. In particular, we ask how long a record survives before being broken, that is, we focus on the lifetimes of records and compile a histogram. Not only is that histogram universal, but it is also exceedingly simple: records lasting m steps occur with a decreasing frequency of $1/m$. We also inquire into some details of these strings. As far as records are concerned, the sequence of numbers can be reduced to a binary string of R's and L's. At each step, we only keep track of whether the record is broken (R) or not (L). A ''tree'' representing all possible sequences can be drawn which displays whether a record is broken or not at each step. We ask: What is the probability that records will be broken at specific steps along the string? The answer, given by Eq. (39) or (32), is also universal. Perhaps most surprising is the answer to the following question: What is the probability that the overall record will occur at the mth step? The answer is *independent* of m. The overall record occurs at any step with equal probability.

Another interesting question is the following. Averaged over many sequences, how does the overall record ''inch up'' with time? Although the answer is not completely universal, the average behavior falls into one of several classes. The simplest is due for a $p(x)$ which is bounded. Not surprisingly, the average just runs into this bound. For unbounded p's, the asymptotic behavior is mostly dictated by the tail of the distribution. In this sense, there is a limited form of universality, that is, properties are independent of the details of the rest of $p(x)$. Beyond the average record, we also studied the probability density, $P(R,t)$, for finding the record R after t steps. Like the central limit theorem, there is universality for large values of t, provided an appropriate time-dependent rescaling of the difference $R-\langle R\rangle$ is included. However, there are three limiting distributions, as well as the possibility of no limiting distribution. These fascinating properties lie outside the scope of this article.

Looking beyond our simple model for ''weather'' records, we may inquire about a natural generalization—a model for ''global warming.'' Here, we let the underlying distribution drift upwards with time. The simplest is a uniform drift, that is, the model distribution at time step t, $p_t(x)$, assumes the form $p(x-\alpha t)$ with $\alpha>0$. Although numerous results still hold, there are also significant differences. Obviously, the average record must increase linearly. Probably the most significant difference between the simple model for ''weather'' records and this ''global warming'' case is that the limiting distribution for $P(R,t)$ is expected to assume the form $P^*(R-\alpha t)$ with no time-dependent rescaling. We are not aware of any general results for such drifting p's. Certainly,

these expectations are confirmed for a flat, drifting $p_t(x)$. For large times, we find $P(R,t)=\partial_R Q(R,t)$, with

$$Q(R,t)=Q^*(R-\alpha t), \tag{35}$$

where

$$Q^*(\xi)=\alpha^k \frac{\Gamma(\xi/\alpha+1)}{\Gamma(\xi/\alpha+1-k)}$$
$$\text{for } 1-k\alpha \leqslant \xi \leqslant 1-(k-1)\alpha. \tag{36}$$

To arrive at these results, we relied on a generalized version of Eq. (15): $Q(R,t)=\Pi_{n=1}^{t} q_n(R)$, where $q_n(R) \equiv \int^R p_n(x)dx$. Clearly, these general forms are applicable for *any* time-dependent p. For example, we could consider a sequence composed of the sum of all random numbers generated before. If the random numbers are distributed such that both positive and negative values are present, we may think of the sequence as the value of a stock, making gains and losses from day to day. Unlike the simple model presented above, there are very few known results for the distributions of the record lifetimes in general. Exploring the universality classes would be a task both daunting and rewarding.

ACKNOWLEDGMENTS

We are grateful for discussions with S. Redner. This research is supported in part by grants from the National Science Foundation through the Division of Materials Research.

APPENDIX A: THE INTEGRAL FOR EQ. (27)

To obtain the desired result Eq. (27), we start from Eq. (17) and the inverse of Eq. (24),

$$t\int_0^1 dq[-\ln(1-q)]q^{t-1}=t\int\left[\sum_{n=1}^{\infty} q^n/n\right]q^{t-1}$$
$$=\sum_{n=1}^{\infty}\frac{t}{n(n+t)}$$
$$=\sum_{n=1}^{\infty}\left[\frac{1}{n}-\frac{1}{n+t}\right]=\sum_{k=1}^{t}\frac{1}{k}. \tag{37}$$

Those who like a bit more mathematical rigor may consider $\lim_{s\to 1}\int_0^s dq\ldots$, which justifies the exchange of the integral and sum.

APPENDIX B: GENERAL STRING PROBABILITIES

We provide some details for computing the general case $P_N(\tau_1,\tau_2,\ldots,\tau_k)$. First, let us introduce an alternate way of labelling the vertices. Instead of the record lifetimes $\{\tau_1,\tau_2,\ldots,\tau_k\}$, we keep track of the times $\{r_1,r_2,\ldots,r_k\}$ when records are broken: that is, r_i denotes the position of the ith R along the string. So, we may also denote the general string probability by $P_N(r_1,r_2,\ldots r_k)$. By convention $r_1=1$, while $r_2=r_1+\tau_1$, $r_3=r_2+\tau_2$, etc.: $\tau_1=r_2-r_1=r_2-1$, $\tau_2=r_3-r_2,\ldots\tau_{k-1}=r_k-r_{k-1}$, $\tau_k=N+1-r_k$, where the last relation comes from the length of the string: $N=r_k+\tau_k-1$.

To obtain P for an arbitrary string, we start with

$$P_N(r_1,r_2,\ldots,r_k)=\int dq_1\cdots dq_N\times\Theta\cdots\Theta. \tag{38}$$

To say that the first record R_1 associated with q_1 has a lifetime of τ_1 means that the next (τ_1-1) R's (and their associated q's) are lower. Therefore, the first (τ_1-1) factors in the integrand are $\Theta(q_1-q_2)\Theta(q_1-q_3)\cdots\Theta(q_1-q_{r_2-1})$, so that the integration over the variables $q_2,q_3,\ldots,q_{r_2-1}$ can be performed trivially, with the result $q_1^{r_2-2}$.

Now, the next record is R_{r_2} associated with q_{r_2}, so that the next factor in the integrand must be $\Theta(q_{r_2}-q_1)$. There will be no more appearances of q_1 in the rest of the integrand. So, the integral over q_1 (up to q_{r_2}) can be performed, giving $q_{r_2}^{r_2-1}/(r_2-1)$. Note that $\tau_1=r_2-r_1=r_2-1$, so that this result can be written simply as $q_{r_2}^{r_2-1}/\tau_1$.

In a similar way, we integrate over the r_3-r_2-1 variables $q_{r_2+1},\ldots,q_{r_3-1}$ up to q_{r_2}, arriving at $q_{r_2}^{r_3-2}/\tau_1$ or $q_{r_2}^{r_3-2}/(r_2-1)$. The integral over q_{r_2} gives $q_{r_3}^{r_3-1}/(r_2-1)\times(r_3-1)$. This process can be carried on until the last integration (over q_{r_k}, up to 1). The integrand consists of the result from the integrals over the previous variables $q_{r_k}^{r_k-1}/(r_2-1)(r_3-1)\cdots(r_k-1)$ as well as the q's from the rest of the sequence: $q_{r_k+1},\ldots,q_N$. With this additional factor of $q_{r_k}^{N-r_k}$, the last integral provides a factor of N. The final result is

$$P_N(r_1,r_2,\ldots,r_k)=\frac{1}{(r_2-1)(r_3-1)\cdots(r_k-1)N}. \tag{39}$$

Rewriting the r's in terms of the τ's, we have

$$P_N(\{\tau\})=\frac{1}{\tau_1(\tau_1+\tau_2)\cdots(\Sigma_1^k\tau_i)}, \tag{40}$$

where the last factor is precisely N, because the sum of the lifetimes of all records is just the length of the string.

One interesting property of these P's follows from the fact that, given a particular string of length N, we can ''generate'' two strings of length $N+1$, by concatenating either an R or an L at the end. In the former case, τ_k stays the same while $\tau_{k+1}=1$. In the latter case, the value of τ_k increases by 1. From Eq. (32), we obtain the following ''recursion'' relations:

$$P_{N+1}(\tau_1,\tau_2,\ldots,\tau_k+1)=\frac{N}{N+1}P_N(\tau_1,\tau_2,\ldots,\tau_k), \tag{41}$$

$$P_{N+1}(\tau_1,\tau_2,\ldots,\tau_k,\tau_{k+1}=1)=\frac{1}{N+1}P_N(\tau_1,\tau_2,\ldots,\tau_k). \tag{42}$$

Not surprisingly,

$$P_{N+1}(\tau_1,\tau_2,\ldots,\tau_k,1)+P_{N+1}(\tau_1,\tau_2,\ldots,\tau_k+1)$$
$$=P_N(\tau_1,\tau_2,\ldots,\tau_k). \tag{43}$$

To illustrate, consider the first pair of entries at the fourth level in Fig. 2 and their relationship to the first entry at the third level $(N=3)$. From Fig. 2(c), we read off $P_3(3)=1/3$, $P_4(4)=1/4$, and $P_4(3,1)=1/12$. These quantities satisfy Eqs. (41) and (42).

APPENDIX C: PROBABILITY FOR STRINGS WHERE THE LAST RECORD SURVIVED m STEPS

An easy way to arrive at the result $T_N(m)=1/m$ for the lifetime histogram is through the probability for the last record to survive m steps, regardless of what happened earlier. Let us define this quantity as $S_N(m)$. To be precise, it is

$$S_N(m)\equiv\sum_{\{\tau\}}\ \delta(\tau_k-m)P_N(\{\tau\}), \tag{44}$$

where δ is the Kronecker delta (δ is unity if its argument vanishes and zero otherwise). The sum notation in Eq. (44) stands for summing over k as well, because k is the *number of records* in the string. Now, because the ''last'' record is also the overall record, that is, all other Rs are lower, it is easy to calculate $S_N(m)$. Because we are summing over all possible ways that the records before this step ($l=N-m+1$) are broken, we may write

$$S_N(m)=\int dq_1\ldots dq_N\times\Theta(q_l-q_1)\ldots\Theta(q_l-q_{l-1})\Theta(q_l-q_{l+1})\ldots\Theta(q_l-q_N)=\int dq_l(q_l)^{N-1}=1/N. \tag{45}$$

As remarked in the main text, this result is not intuitively obvious. Intuition might lead us to an $S_N(m)$ with a maximum in the middle of the string, but Eq. (45) tells us that the distribution is flat.

APPENDIX D: HISTOGRAM FOR RECORD LIFETIMES $T_N(m)$

Note that $T_N(m)$ is not a real probability in the sense that $\Sigma_m T\geqslant 1$. The normalization condition is

$$\sum_{\{\tau\}}\ P_N(\{\tau\})=1, \tag{46}$$

but the histogram is defined by

$$T_N(m)=\sum_{\{\tau\}}\ P_N(\{\tau\})\sum_{i=1}^{k}\ \delta(\tau_i-m). \tag{47}$$

Again, we remind the reader that the $\Sigma_{\{\tau\}}$ involves a sum over k as well. Summing over m produces $\Sigma_{\{\tau\}}P_N k\geqslant\Sigma_{\{\tau\}}P_N=1$.

Now, given a string of N random numbers, the only way for a record lifetime to be N is that the first record survives. Thus, $T_N(N)$ is precisely $P_N(\tau_1=N)$, so that

$$T_N(N)=1/N. \tag{48}$$

Our goal reduces to proving

$$T_{N+1}(m)=T_N(m)\quad\text{for}\quad 1\leqslant m<N. \tag{49}$$

There are two types of contributions to $T_N(m)$. One is from all the ''interior'' records:

$$\tilde{T}_N(m)=\sum_{\{\tau\}}\ P_N(\{\tau\})\sum_{i=1}^{k-1}\ \delta(\tau_i-m), \tag{50}$$

that is, $\tau_i=m$ with $i<k$. The other piece, from the last record alone, is precisely $S_N(m)$. For $\tilde{T}_{N+1}$, let us start with $\tilde{T}_N(m)$ and go from N to $N+1$ by looking at

$$\sum_{\{\tau\}}\ P_{N+1}(\{\tau\})\sum_{i=1}^{k-1}\ \delta(\tau_i-m), \tag{51}$$

where the ks appearing here are still those associated with $\tilde{T}_N(m)$, that is, the last record may be labeled by $k+1$. Because the sum over τ_k and τ_{k+1} can be performed without the δ's, Eqs. (41) and (42) can be used to obtain the following recursion relation:

$$\sum_{\{\tau\}}\ P_{N+1}(\{\tau\})\sum_{i=1}^{k-1}\ \delta(\tau_i-m)=\sum_{\{\tau\}}\ P_N(\{\tau\})\sum_{i=1}^{k-1}\ \delta(\tau_i-m). \tag{52}$$

But the quantity in Eq. (51) is not the only contribution to $\tilde{T}_{N+1}$, because concatenating an R to an N string will produce an interior record for the $(N+1)$ string. Focusing on a specific m, this extra bit is just

$$\sum_{\{\tau\}}\ P_{N+1}(\tau_1,\tau_2,\ldots,\tau_k,\tau_{k+1}=1)\,\delta(\tau_k-m), \tag{53}$$

which is

$$\sum_{\{\tau\}}\ \frac{1}{N+1}P_N(\tau_1,\tau_2,\ldots,\tau_k)\,\delta(\tau_k-m)=\frac{1}{N+1}S_N(m). \tag{54}$$

Hence, we conclude

$$\tilde{T}_{N+1}(m)=\tilde{T}_N(m)+\frac{S_N(m)}{N+1}. \tag{55}$$

With the help of Eq. (55), we have

$$T_{N+1}(m)=\tilde{T}_{N+1}(m)+S_{N+1}(m)=\tilde{T}_N(m)+\frac{S_N(m)}{N+1}+S_{N+1}(m). \tag{56}$$

But, according to Eq. (45),

$$S_{N+1}(m)=\frac{N}{N+1}S_N(m), \tag{57}$$

so that

$$\tilde{T}_{N+1}(m)=\tilde{T}_N(m)+S_N(m)=T_N(m). \tag{58}$$

Thus, using (48), we arrive at

$$T_N(m)=1/m \tag{59}$$

for any $N\geqslant m$.

[1]The data can be found at: http://www.wdbj7.com/climate/climate.htm.

[2]See, for example, E. J. Gumbel, *The Statistics of Extremes* (Columbia U.P., New York, 1958); J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Wiley, New York, 1978). Gumbel gives a short historical summary in his book.

[3]See, for example, H. Gould and J. Tobochnik, *An Introduction to Computer Simulation Methods* (Addison-Wesley, Reading, MA, 1996), Sec. 11.5.

[4]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes* (Cambridge U.P., Cambridge, 1992), 2nd ed.

[5]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1974).

[6]Record times are discussed, for example, by Galambos (see Ref. 2).