CORRESPONDENCE

Comments on “Plotting Positions in Extreme Value Analysis”

NICHOLAS COOK
RWDI Anemos, Ltd., Dunstable, United Kingdom

(Manuscript received 9 June 2009, in final form 23 February 2010)

ABSTRACT

This comment addresses the role of sampling error in extreme value analysis. A note published in this journal claimed that Weibull’s 1939 estimator for sample probability has a unique status that invalidates all other estimators and renders invalid all of the developments of unbiased distribution-dependent estimators made since 1939. The note concluded that the use of distribution-dependent estimators should be abandoned and that many estimates of the weather-related risks should be reevaluated and the related building codes and other related regulations updated. This comment uses rigorous statistical proofs to make the diametrically opposite case: namely, that development of distribution-dependent estimators has resulted in an improvement in accuracy over the past half century and that no changes are required to the basis of weather-related building codes and regulations. These rigorous proofs are supplemented by sampling experiments that demonstrate their validity. This comment provides an introduction to the basic statistical concepts of the statistical modeling of extremes, including unbiased estimators for the model parameters.

1. Introduction

In a note published in this journal (Makkonen 2006, hereinafter referred to as “the note”) posed the question that in “order-ranked data by which the return periods of specific extreme events are estimated” . . . “What is the cumulative probability \( P \) that should be associated with the sample of rank \( m \)?” The note claimed that the estimator \( P = m/(N + 1) \) derived by Weibull (1939) should be used exclusively to derive all statistical properties, including mean recurrence interval, or return period, and also the plotting positions on Gumbel axes. The note declared that the improvements in extreme value analysis methods developed since 1939 are “invalid” and that the “many estimates of the weather-related risks should be reevaluated and the related building codes and other related regulations updated.” Publication of the note resulted in a press release by the American Meteorological Society under the title “Natural hazards are more common than statistics indicate.”

If valid, the note should have provoked an immediate and urgent response. The only published response in this journal has been the comments by de Haan (2007) that the note “addresses a technical problem of limited importance” since it “is concerned with the use of Gumbel probability paper” and fitting a Gumbel distribution “can be done much more efficiently by estimating its parameters (e.g., by maximum likelihood methods) than by plotting a line on Gumbel paper.” De Haan did not challenge the validity of the note but was more concerned with the tone and content of the associated press release. One issue raised by de Haan is particularly relevant: “the purpose of extreme value analysis is not discussed in the paper” [emphasis added by this author].

Many parameter-fitting methods that do not require direct estimates of sample probability in their application nevertheless incorporate assumptions into their derivation, and therefore the issue is of more importance than de Haan suggests. In any case, there are situations in which the ability to plot an unbiased sample distribution and fit this distribution to a model distribution using the weighted least mean squares method is useful—in particular, when fitting observations of wind speeds from mixed climates to a composite distribution (e.g., Cook et al. 2003; Cook 2004) or when comparing...
penultimate distributions with their asymptotes (e.g., Cook and Harris 2004, 2008).

The purpose of this comment is to demonstrate, using rigorous proofs and demonstrations, that the unbiased estimators developed since 1939 are valid and provide a significant improvement in accuracy. This cannot be done without directly addressing the role of sampling error.

2. Probability

First, to help to understand the role of sampling error, it is instructive to review the meaning of “random” and the concept of “probability.” Cramer’s view (Cramer 1946, p. 138) is that random cannot be precisely defined and that its sense is best described by reference to real-world examples.

Cramer uses the classic coin-tossing experiment (Cramer 1946, 142–143) to demonstrate that, if “heads” occurs \( n \) times in \( n \) tosses of the coin, the frequency ratio \( n/n \) “shows a marked tendency to become more or less constant for large values of \( n \)” and that after a long series of repetitions the frequency would approximate to the probability \( P \). Cramer describes this as “the frequency interpretation of the probability \( P \).”

Cramer’s coin-tossing experiment was performed 400 times (Cramer 1946, p. 143), after which the mean frequency ratio is still not 0.5. With the advent of personal computers with built-in pseudorandom number generators and of efficient algorithms to eliminate any residual serial correlation in the pseudorandom numbers (Press et al. 1989), very many trials of more complex random experiments can be performed easily.

Other authors, for example, von Mises (1941), apply an axiomatic definition for probability—as the limit of the frequency ratio \( v/n \) as the number of trials approaches infinity, or

\[
P = \lim_{n \to \infty} (v/n). \tag{1}\]

The accumulation of a mean value is not a limiting process in the conventional sense, because the difference between each successive trial does not diminish, and therefore this simple definition can cause difficulties. As an alternative, for compatibility with analysis of extremes, the \( n \) trials can be regarded as being made up of \( K \) trials each of \( N \) observations, such that \( KN = n \). Then

\[
P = \lim_{K \to \infty} \frac{1}{K} \sum_{i=1}^{K} \frac{v_i}{N} = \frac{P}{N}. \tag{2}\]

Sampling experiments, which are often referred to as “bootstrapping” (Efron and Tibshirani 1993), exactly replicate this frequency interpretation of \( P \) (except, for practical reasons, \( K \) is large but finite) and are therefore a convenient way to demonstrate sampling theory and the role of sampling error. Table 1 shows the results of a simple sampling experiment: 10 trials, each of \( N = 10 000 \) samples, were obtained from a conventional probability generator. The frequency was counted of the values below the thresholds 0.1–0.9 in 0.1 increments. It is clear that, although the results are close to the expected frequencies, every trial is slightly different because of the action of random sampling error.

3. Sampling error

Section 2 established that the concept of probability is precise only in the context of an infinite number of trials. When a real-world observation is made, it is done in the expectation that an ideal invariant process has been

---

1 Because of space restrictions, only the starting expression and final result of each proof is given in this paper. The complete proofs were submitted for review with the paper and are available on application by e-mail to the author: nick.cook@rwdi.com or njcook@ntlworld.com.

2 One that provided uncorrelated random numbers in the range 0–1 drawn from a rectangular distribution.
sampled. Thus all of the properties of this ideal process can be regarded as fixed, precise values: including probability \( P_t \), risk of exceedance \( Q_t \), and recurrence interval \( T_t \).

A value associated with any one trial sample is an imprecise estimate, however, and is subject to sampling error. Sampling error \( e \) has two components: 1) a systematic mean component—the bias error \( \bar{e} \)—and 2) a random component with zero mean—the variance error—with the standard deviation \( \sigma(e) \). It is possible for there to be no bias (\( \bar{e} = 0 \)), but the variance error \( \sigma(e) \) is greater than 0 for all finite-length samples. Most, if not all, extreme value analysis theory addresses this sampling error in such a way that bias error is eliminated and variance error is minimized so that the “best” estimate is obtained. Hence, any real-world situation corresponds to a single trial in which the sample values of the statistical parameters differ from the ideal values by the sample error: sample probability \( P = P_t + e_P \), sample risk \( Q = Q_t + e_Q \), and recurrence interval \( T = T_t + e_T \).

Note also that, in any real-world situation, there will also be uncertainty in the measurement of the value: the observational error. Measurements would typically be expected to be accurate to \(<1\% \) but, in the case of wind speed, the World Meteorological Organization standard procedure requires the value to be rounded to the nearest knot \((\sim0.5 \text{ m s}^{-1})\). This requirement results in an observational error \( e_v \) with standard deviation \( \sigma(e_v) \approx 0.15 \text{ m s}^{-1} \). As this value is small in comparison with the sampling error, the sampling error dominates.\(^3\) Observational error is not considered in this paper.

### 4. Return period

“Return period” is conventionally defined as being the mean recurrence interval \( T \) (e.g., Gumbel 1958, p. 23; Castillo 1988, p. 15), and for the ideal invariant process

\[
R_t = T_t = 1/Q_t = 1/(1 - P_t). \quad (3)
\]

The sample return period, that is, the return period associated with a value \( x \) of a finite sample, can be defined as

\[
R(x) = 1/Q(x) = 1/[1 - P(x)]. \quad (4)
\]

Equation (4) is commonly used as a transformation to express the probability or risk of a time-varying process in units of time, a measure that is more familiar to engineers. Because this expression for \( R(x) \) is nonlinear in both \( P(x) \) and \( Q(x) \), it has implications for the concomitance of expressions containing means of a sample value. Hence, from Eqs. (2) and (4), the mean recurrence interval

\[
\overline{R}(x) = T(x) = \frac{1}{\overline{Q}(x)} = \frac{1}{1 - \overline{P}(x)} \rightarrow R_t \quad \text{as} \quad n \rightarrow \infty. \quad (5)
\]

Table 2 lists the values from \( n = 10^6 \) trials (accurate to \( \sim0.1\% \)) at the same thresholds as Table 1 for the mean sample probability \( \overline{P} \); its complement, the risk of exceedance \( \overline{Q} = 1 - \overline{P} \); the reciprocal of \( \overline{Q} \); and the mean recurrence interval \( \overline{T} = 1/(1 - \overline{P}) \). It is seen that the values of \( \overline{P} \) and \( \overline{Q} \) in each column of Table 2 always sum to unity: they are concomitant because they are linearly related; \( 1/\overline{Q} \) and \( \overline{T} \) are not equal, however, because the mean of the reciprocal is not the reciprocal of the mean:

\[
\overline{T}(x) \neq 1/\overline{Q}(x) \quad \text{for any finite} \quad n. \quad (6)
\]

The consequences of this inequality are examined in more detail later.

### 5. Order statistics

The analysis of extreme wind speeds deals with sets of samples and operates through their order statistics, that is, their rank in order of value. The convention is to assign a rank \( m \) from \( m = 1 \) to \( m = N \), in ascending order of value—although descending order is sometimes used when this simplifies expressions (especially those framed in terms of \( Q \)). Here the subscript \( N \) is used to denote the \( m \)th ranked value out of \( N \), where \( m = 1 \) is the smallest and \( m = N \) is the largest.

Figure 1 shows the mean probability \( \overline{P}_{m,N} \) accumulating for the \( m = 1, \ldots, 9 \) ranked samples of size \( N = 9 \).

### Table 2. Mean probabilities, risks, and recurrence intervals below various threshold probabilities, for \( n = 1 \,000 \,000 \) trials.

<table>
<thead>
<tr>
<th>Threshold ( P )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \overline{P} = \frac{n}{n} )</td>
<td>0.0998</td>
<td>0.1994</td>
<td>0.2998</td>
<td>0.3994</td>
<td>0.5000</td>
<td>0.6002</td>
<td>0.7004</td>
<td>0.8008</td>
<td>0.9005</td>
</tr>
<tr>
<td>( \overline{Q} = \frac{n - P}{n} )</td>
<td>0.9002</td>
<td>0.8006</td>
<td>0.7002</td>
<td>0.6006</td>
<td>0.5001</td>
<td>0.4009</td>
<td>0.3007</td>
<td>0.2003</td>
<td>0.1001</td>
</tr>
<tr>
<td>( 1/\overline{Q} )</td>
<td>1.1110</td>
<td>1.2490</td>
<td>1.4280</td>
<td>1.6650</td>
<td>2.0000</td>
<td>2.5010</td>
<td>3.3380</td>
<td>5.0190</td>
<td>10.050</td>
</tr>
<tr>
<td>( \overline{T} )</td>
<td>0.1108</td>
<td>0.2491</td>
<td>0.4281</td>
<td>0.6651</td>
<td>2.3380</td>
<td>4.0190</td>
<td>9.0480</td>
<td>9.0480</td>
<td>9.0480</td>
</tr>
</tbody>
</table>

\(^3\) To be in accord with this condition, Castillo (1988) and Harris (1996) recommend that, in regressions, probability be taken as the dependent variable and wind speed as the independent variable, which is the reverse of the usual convention.
as the number of trials \( n \) increases. The value \( N = 9 \) was used, for convenience, to demonstrate that \( P_{m,N} \) stabilizes toward \( m/10 \). Note that this stabilization is very slow: after 1000 averages the error from the expected value is typically 0.005, and even after 10 000 averages the error is greater than 0.001. Figure 1 demonstrates that, in the real-world example of annual maxima from a record period of 9 yr, 10 000 nine-year periods would be required to estimate a value of \( P \) accurate to 1% and that the estimate of \( P \) from any single 9-yr period is far less precise.

6. Expectations

a. Mean value, or expectation

As shown above, the concept of probability depends on the mean frequency stabilizing over a large number of trials. It follows directly that the best estimate of any random property \( x \) is this mean \( \bar{x} \) for a large number of trials. Provided that the trials meet the requirements of statistical independence, recurrences of events are given by the binomial distribution, and this may be used to evaluate the “expectation,” the theoretical mean over an infinite number of trials.

In this paper, the notation \( E(x) \) is used to denote the expectation, that is, the mean value derived from theory, and the notation \( \bar{x} \) is used to denote the sample mean from the observations or experiment. Also the standard notation \( F[G(x)] \) is used to denote the transformation \( F() \) applied to the value of \( G(x) \), as in \( P[E(R)] \) for the probability associated with the expectation of \( R \). Further, because \( P_{l}(x) \) is always a single-valued function of \( x \), any function \( G(x) \) can be expressed instead as a function of \( P_{l} G(P_{l}) \).

The expectation of the \( m \)th ranked value of any transformation \( G(P_{l}) \) of the variable \( x \) is defined as

\[
E(G_{m:N}) = \frac{N!}{(m-1)!(N-m)!} \int_{0}^{1} G(P_{l}) P_{l}^{m-1} (1 - P_{l})^{N-m} dP_{l} \tag{7}
\]

\{Gumbel 1958, Eq. [2.1.1(1)]; Castillo 1988, Eq. (2.3)\}

This is easily evaluated and leads to Weibull’s estimator (Weibull 1939):

\[
E(P_{m:N}) = \frac{m}{N+1} \tag{9}
\]

\{Gumbel 1958, Eq. [2.1.3(4)]; Castillo 1988, Eq. (2.41)\}

\( m \) is the uncertainty associated with these values is given by the standard deviation,

\[
\sigma = \sqrt{\frac{1}{N-1} \int_{0}^{1} (G(P_{l}) - E(G_{m:N}))^2 dP_{l}}
\]

\{Gumbel 1958, Eq. [2.1.3(4)]; Castillo 1988, Eq. (2.41)\}
The standard deviations, \( s(P_m: N) \), for the experiment of Fig. 1 are compared with the expectations from Eq. (10) in Fig. 2. This shows that the 10,000 trials reproduce the expected behavior extremely well.

**c. Expectation of risk and its standard deviation**

By repeating the above principle, the best estimate of the sample risk of exceedance is its expectation \( E(Q_m: N) \) obtained by substituting \( G(P) = 1 - P \) in Eq. (7), leading to

\[
E(Q_m: N) = \frac{N - m}{N + 1}. 
\]

Equation (13) assigns an infinite return period to the highest-rank \( m = N \) because the integrand has a singularity there. The expectation includes the maximum possible return period (for \( P = 1 \)), which is infinite, and the highest-ranked sample return period is therefore "unplottable." The probability corresponding to \( E(R) \) is therefore given by

\[
P(E(R_m: N)) = \frac{1}{N}. 
\]

The standard deviations \( \sigma(P_m: N) \) for the experiment of Fig. 1 are compared with the expectations from Eq. (13) in Fig. 2. This shows that the 10,000 trials reproduce the expected behavior extremely well.

**d. Expectation of return period and its standard deviation**

The expectation \( E(R_m: N) \) is obtained by substituting the sample risk of exceedance from Eq. (7) leading to

\[
E(R_m: N) = \frac{1}{N - m}. 
\]

Equation (13) assigns an infinite return period to the highest-rank \( m = N \) because the integrand has a singularity there. The expectation includes the maximum possible return period (for \( P = 1 \)), which is infinite, and the highest-ranked sample return period is therefore "unplottable." The probability corresponding to \( E(R) \) is therefore given by

\[
P(E(R_m: N)) = \frac{1}{N}. 
\]
The uncertainty associated with $E(R_{m:N})$ is given by the standard deviation $\sigma(R_{m:N})$:

$$
\sigma(R_{m:N}) = \left[ \frac{mN}{(N-m)^2(N-m-1)} \right]^{1/2}.
$$

Note that the standard deviations of both the highest and second-highest ranks, $m = N$ and $m = N - 1$, are infinite. So, although the second-highest ranked value can be plotted, the infinite variance error indicates that it should be given zero weight in any model fit. In fitting extreme value data directly to the return period, the highest and second highest values serve only to set the plotting positions of the lower ranks. In the 9 yr of data represented here, only seven points contribute to the fit, indicating that return period is an extremely poor choice of fitting parameter.

Figure 5 shows the corresponding standard deviations for the samples in the experiment of Fig. 1 in comparison with the expectations $E(R_{m:N})$ from Eq. (15). The match is good where the expectations remain finite.

e. Expectation of the Fisher–Tippet type-1 variate $y$ and its variance

For the extremes of a physical variate, such as wind speed $V$ or, more significantly, wind actions (pressures, loads, and moments, proportional to $V^2$), the best estimate is again the expectation of the physical variate. The Fisher–Tippet type-1 (FT1) distribution is frequently adopted as the model distribution for wind actions and is defined by

$$
P(x) = 1 - \exp\{-\exp[a(x - U)]\}.
$$

Here, $x$ is the variate, $U$ is its mode, and $1/a$ is its dispersion. The current hotly debated issues of convergence, generalized extreme value, and penultimate FT1 distributions are left to another forum, and the FT1 distribution is taken as the exemplar here. It is conventional to break Eq. (16) down into its two constituent parts:

$$
y = a(x - U) \quad \text{and} \quad (17)
y = -\ln[-\ln(P)]. \quad (18)
$$

The parameter $y$ in Eqs. (17)–(18) is Gumbel’s “reduced variate” \{Gumbel 1958, Eq. [1.1.1(1)]\} and makes the variate nondimensional. Hence, in Eq. (7) the required function is $G(P_i) = -\ln[-\ln(P_i)]$, and this function is not amenable to a simple analytical evaluation, but it can be evaluated numerically. Gumbel includes a chart of $E(y_{m:N})$ for various values of $N$ calculated laboriously by Kimball (Gumbel 1958, p. 211) using an International Business Machines, Inc., tabulating machine as a difference engine. The various alternative plotting positions for extremes suggested over the years have arisen because it was not possible to evaluate the integrals for $E(y_{m:N})$ as...
part of an analysis until inexpensive and fast computing became available (e.g., Harris 1996). The estimator of Gringorten (1963), \( P = (m - 0.44)/(N + 0.12) \), becomes

\[
E(y_{m,N}) = -\ln[-\ln(m - 0.44) + \ln(N + 0.12)]
\]

(19)

and matches the FT1 distribution very well—a fact that is easily verified by sampling the distribution of \( y \) and accumulating \( y_{m,N} \) over a very large number of trials.

Gumbel (1958, Eq. [6.1.1.(2)]) , however, gives the analytical solution for the highest value: \( E(y_{y,N}) = \gamma + \ln(N) \), where \( \gamma = 0.5772157 \) is Euler’s constant, which for \( N = 9 \) gives \( E(y_{y,N}) = 2.774 \). Gringorten’s estimator gives \( E(y_{y,N}) \approx 2.759 \), whereas the bootstrap with \( n = 10\,000 \) trials gives \( E(y_{y,N}) \approx 2.788 \), showing that both Gringorten (1963) and the numerical approach give excellent estimates.

---

**Fig. 4.** Mean return period for each rank of ranked samples of size \( N = 9 \); open circles denote samples from \( n = 10\,000 \) trials, plus signs indicate the expectation, and the times signs indicate derived from mean risk. The samples match the expectation, except where the expectation is infinite. Return periods derived from the mean risk are underestimates and therefore exaggerate design values.

**Fig. 5.** Std dev of return period for each rank of ranked samples of size \( N = 9 \); open circles denote samples from \( n = 10\,000 \) trials and plus signs indicate the expectation. The samples match the expectations, except for the top two ranks, for which the expectations are infinite.
For the standard deviation $\sigma(y_{m:N})$ there is again no easy alternative available to a numerical computer-based approach. The values of $E(y_{m:N})$ and $\sigma(y_{m:N})$ obtained by numerical integration methods form the core of the procedure derived by Harris (1996), giving the plotting positions and fitting weights for a weighted least squares regression.

Figure 6 shows the mean reduced variate $\gamma$ plotted against the various estimates of probability on conventional Gumbel axes. The ranked values of $P$ from each of the $n = 10,000$ trials from the sampling experiment of Fig. 1 were transformed to $-\ln(-\ln(P))$, giving ranked values of $\gamma$ drawn from an FT1 distribution, before averaging each rank. The results were fitted back to an FT1 distribution by a least mean squares fit, and the parameters are tabulated in Table 3. In addition, results are given for the best linear unbiased estimator (BLUE) fitting method of Lieblein (1974), which produces an unbiased fit directly from the ranked values without recourse to the associated probability estimate.

The expectation for a perfect fit in Fig. 6 is a slope of unity and an intercept of 0. The Weibull plotting positions from $E(P)$ give a good estimate of the intercept but overestimate the slope by 18%, and therefore the datum design value of $\gamma$ for $P = 0.98$ ($R = 50$) is overestimated by 18%. The plotting positions from $E(R)$ underestimate both slope and intercept, and therefore the datum design value of $\gamma$ for $R = 50$ is underestimated by 12.5%. The plotting positions of Gringorten (1963) give a very small underestimate of slope and a very small positive intercept, and therefore the datum design value of $\gamma$ is overestimated by only 0.08%. The numerical method of Harris (1996) gives a very small overestimate of slope and intercept, and therefore the datum design value of $\gamma$ is overestimated by 0.65%. The method of Lieblein (1974) gives an almost exact fit, and the datum design value is accurate to 0.01%.

**7. Review of the case made in the note**

**a. “Introduction”**

The note (Makkonen 2006) claims in its introduction section that return period is the primary datum parameter for engineering design to resist wind loads. This may have been true 40 years ago, when the first statistically
Table 3. Model parameters fitted to the mean of n = 10 000 trials of ranked samples of size N = 9 drawn from an FT1 distribution.

<table>
<thead>
<tr>
<th>Basis of fit</th>
<th>Slope</th>
<th>Intercept</th>
<th>$y(P = 0.98)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(P) = m(n + 1)$</td>
<td>1.1818</td>
<td>-0.00259</td>
<td>4.609 (+18%)</td>
</tr>
<tr>
<td>$P[E(R)] = m/N$</td>
<td>0.9102</td>
<td>-0.1386</td>
<td>3.413 (-12.5%)</td>
</tr>
<tr>
<td>Gringorten (1963)</td>
<td>0.9894</td>
<td>0.0447</td>
<td>3.905 (+0.08%)</td>
</tr>
<tr>
<td>Lieblein (1974)</td>
<td>0.9998</td>
<td>0.00036</td>
<td>3.902 (-0.01%)</td>
</tr>
<tr>
<td>Harris (1996)</td>
<td>1.0021</td>
<td>0.0021</td>
<td>3.912 (+0.26%)</td>
</tr>
</tbody>
</table>

based wind code was published (BSI 1970), but is not the case today. The problem with using return period as the reference datum is that it was interpreted by many engineers as being equivalent to a “design lifetime.” A deliberate change to the description of the primary datum parameter for buildings from $R = 50$ yr to annual risk of exceedance $Q = 0.02$ was introduced into many design codes in the 1990s (e.g., BSI 1995) to emphasize that the corresponding risk of failure applied to each and every year the building is exposed to the wind, irrespective of the building’s design lifetime.

Most current codes of practice for buildings, bridges, and other engineering structures now use a reliability-based approach to setting the design wind loads, the design resistance of the structure, and the corresponding safety factors. This is done by ensuring that wind load times $g_f$ is less than structural resistance divided by $g_m$, where $g_f$ and $g_m$ are partial safety factors for load and material strength, respectively, with values set to control the probability of failure (expressed as a “reliability index” $\beta$). From this, it can be seen that the design wind load for the datum risk is not the sole parameter; it is the upper tail of the distribution that sets the reliability index $\beta$ that is important. Hence the emphasis in assessing wind speed observations should not be focused on getting the best value for just the datum risk, but should be focused on getting the best estimate of the upper tail of the distribution and hence of both FT1 parameters: mode and dispersion.

b. “The history of plotting positions”

In this titled section, the note claims that the Weibull estimator $P = m/(N + 1)$ plays a unique role. Authors that have used other estimators, for example, modal or median values, are quoted in a manner that suggests that they may have confused and contradictory views on the subject. As noted by de Haan (2007), the note does not address the differing purposes for making these estimates. When the purpose is to obtain the most likely value, the modal estimate is appropriate. If the issue is when a value may next be exceeded, this is equally likely to occur before or after the median return period. Most often, the purpose is to fit a model distribution using methods that require the expectation.

Four sources are cited in the note to support the claim for the unique role of the Weibull estimator in extreme value analysis.

1) GUMBEL (1958)

Gumbel indeed advocates the use of $P = m/(N + 1)$ but makes it clear that this produces modal plotting positions $\hat{y}$, that is, values that correspond to the mode of the sampling error. Gumbel denotes the corresponding modal recurrence interval by $T(\hat{y})$, for example, in his Table 6.1.2(2), to distinguish this from the mean recurrence interval $T$. Plotting the mode $\hat{y}$ on Gumbel axes creates a bias error for any fitting method requiring the mean $\bar{y}$. For the highest rank $y_{N,N}$ the error {Gumbel 1958, Eq. [6.1.1.(2)]} is $\pi(\hat{y}_{N,N}) = \gamma = 0.577 215 7$, which is very significant.

2) COOK (1982, 1985)

This paper and text book specifically address the bias error caused by the modal plotting positions $\hat{y}$ and advocate the use of the Lieblein BLUE method (Lieblein 1974) to eliminate this bias.

3) COOK ET AL. (2003)

This paper did not use the Weibull estimator at all but used the mean plotting positions for $\bar{y}$ from Harris (1996), derived numerically.

4) SUMMARY

It is therefore clear that none of these four citations support the argument of a unique role for the Weibull estimator, and their citation is perverse.

c. “The correct probability positions for estimating return periods”

The first step in this section of the note is to derive the expectation $E(P)$ from the binomial distribution, resulting in the Weibull estimator. Equations (4) and (5) of the note are identical to Eqs. (8) and (9) in this paper—it is not disputed that the Weibull estimator gives the expectation $E(P)$. The second step in the note is to derive the mean rate of exceedance $\bar{T}$ from $n$ trials. In the notation of this paper, $\bar{T} = n \times E(Q)$. Equation (8) in the note differs from Eq. (11) in this paper only by the factor $n$. The third step in the note is formally to define “return period” $R$ as the mean recurrence interval $T$, again the same as in this paper.

The fourth step in this section of the note contains a fundamental error. Equation (9) of the note equates $R = T$ to $1/Q$, which is not true, as established earlier by Eq. (6) and demonstrated in Table 2. Equation (10) in the note is the reciprocal of Eq. (11) here for $E(Q)$, and...
the resulting estimator, \( R[E(Q)] = \tilde{T} \), is the modal recurrence interval. This error propagates into the fifth step where, in Eq. (12) of the note, it leads to the erroneous claim that the Weibull estimator results from “the exact definition of the return period” [italics in original].

The steps from Eqs. (7) through (12) in the note do the following:

\[
\tau/n = \frac{N - m + 1}{N + 1} = E(Q) \Rightarrow R[E(Q)] = \tilde{T} \\
= R[E(P)] \Rightarrow E(P) = \frac{m}{N + 1}.
\]

The ⇒ symbol indicates a transformation to \( R \) and back again, which has no net effect other than to express the modal recurrence interval \( \tilde{T} \). The steps shown by Eq. (20) never achieve “the exact definition of the return period” \( \tilde{T} \) but only prove that \( E(Q) = 1 - E(P) \), which is self-evident.

The conclusion of this section, that the Weibull estimator gives both \( E(P) \) and \( \tilde{T} \), is false: the rigorous proofs, above, show \( E(P) = m/(N + 1) \) [Eq. (9)] and \( P(\tilde{T}) = m/N \) [Eq. (14)]. Both of these results remain “independent of the underlying distribution” [italics in original]. It follows that the claim that the Weibull estimator has a unique role because it is the only distribution-free estimator fails and, in accord with this, so does the assertion that correspondingly all other plotting positions must be incorrect.

d. “Plotting positions involving a reduced variate”

Section 4 of the note makes the following assertions:

(i) The plotting position should always be distribution free.

(ii) Where the variate is transformed, the transformation should always be applied after the plotting position is determined, for example, to give the modal value \( \tilde{y} = y[E(P)] \).

(iii) “[T]here exists a fundamental connection between the rank of an observation and the estimate of its return period” [italics in original].

(iv) The Weibull estimator should always be used: “Eq. (10) must not be manipulated based on an arbitrary choice of the scale on the ordinate axis of the graph that is devised to merely alleviate the analysis of the data.” The “fitting procedure may reflect the scale used, but the probability positions of the data must be the same regardless of the method of fitting. In other words, the plotting formula \( P = m/(N + 1) \) is valid regardless of the transformation made” [italics in original].

(v) Determination of the “plotting positions” and the fitting of observations to a model are two distinct and separate processes.

(vi) The “foundation for the use of the mean \( E[F(x_m)] \) is merely that the return period \( R \) is defined as the mean time period \( T \) between events that exceed \( F(x_m) \)” [italics in original; in the current paper \( F(x_m) = P_{m:N} \)].

(vii) The “linearity, shown by Gumbel (1958) to exist as a result of plotting \( E[F(x_m)] \) by Eq. (3), is lost when \( E(\eta_m) \) is being plotted” [in the current paper \( E(\eta_m) = E(y_m;N) \)].

(viii) The “best estimate of a sample parameter is its mean value only if that parameter is additive” because otherwise “plotting positions no more correspond to the probability \( P \) that is required to estimate the return period.”

(ix) The “concept of distribution-specific plotting formulas in analyzing return periods should be abandoned. This causes no problems to the analysis, however, because the Weibull plotting formula \( P = m/(N + 1) \) is to be used regardless of the underlining distribution.”

In asserting i, Makkonen (2006) follows Gumbel (1958) in desiring a universal distribution-free plotting position. This is an ideal aspiration but is not essential to any pragmatic approach. As a consequence, in ii the Weibull estimator for \( E(P) \) gives the modal plotting position \( \tilde{y} \) on Gumbel axes. Unlike Gumbel, however, the note ignores the bias error this creates, and this bias error depends on the distribution. Statisticians since Gumbel have developed methods to eliminate bias error that are necessarily distribution dependent. Assertion iii is true only for the ideal invariant process: for all finite samples there are a number of estimates each associated with a different purpose, and so assertion iv does not follow. Manipulating data to facilitate analysis is called preconditioning and is a legitimate procedure in common use in many fields. Banning its use without due cause unnecessarily eliminates a large class of valid methods.

The two important issues raised by de Haan (2007) are relevant to assertion v. The note does not address the purpose for extreme value analysis; instead citing authors who declare a specific purpose and use a rigorous axiomatic approach to derive corresponding plotting positions as examples of erroneous practice. The principal purpose of the cited authors is to obtain the best estimates of the parameters of the distribution. It is self-evident that this purpose cannot be achieved on a distribution-free basis. Assignment of the sample probability for each observation is only a precursor to fitting to a distribution, and therefore it is irrelevant to the main purpose whether...
this fit is made in one or two steps. Some methods, such as Lieblein (1974), fit in one step without requiring sample probabilities at all, but these are implicit in the derivation of the methods. For methods that fit using mean values, it makes no sense to adopt modal values and then to have to correct for the resulting bias in a second procedure if the fit can be obtained in one step.

Assertion vi exposes the fundamental error in the note: \( E(P) \) does not lead to the mean time between events (mean recurrence interval \( T \)), a point that was made several times earlier in this paper.

Assertion vii is untrue: the meteorological examples in Gumbel [1958, graphs 6.3.1, 6.3.3(1), and 6.3.3(2)] are forced to a linear fit using \( E(P) \). It is impossible to show that \( E(y) \) produces a less linear fit because \( y \) is linearly related to \( x \) and \( E(x) \) is a linear operation. The Gumbel plot for the sampling experiment, Fig. 6, demonstrates that neither of the distribution-free estimators, \( E(P) \) or \( E(R) \), reproduces the linear form of the datum distribution from which the samples were drawn, but rather they bracket the datum straight line as a convex/concave pair of curves. The distribution-dependent Gringorten (1963) estimator and the Harris (1996) method reproduce the sampled distribution almost exactly.

Assertion viii is not true. It follows from the frequency interpretation of probability that the best estimate of any random variate is its mean value accumulated over a very large number of trials, that is, tends to the expectation. The note’s assertion that the plotting position should always be determined by the Weibull estimator before any linearizing transformations are applied violates this principle. The note asserts that “\( P \) is being estimated,” but this stems from two sources of error: 1) the assumption that the required parameter is the return period \( R \) and 2) the assumption that \( E(R) \) and \( E(P) \) are concomitant, when neither assumption is true. The purpose of the Gumbel axes is to linearize the parameters of the distribution with respect to the variate, and therefore it is \( y \), and not \( R \), that is being estimated. Because \( y(P) \) is also nonlinear, the best estimates of \( P, R \), and \( y \) are not mutually concomitant.

Assertion ix, that all distribution-dependent estimators should be abandoned in favor of the Weibull estimator, fails under the weight of all the contrary evidence presented above. Indeed the opposite is conclusively demonstrated in that only the distribution-dependent methods reproduce the original sampled distribution. Of course, the bias errors reduce in inverse proportion to the square root of the number of observations. The small sample size \( N = 9 \) used here was chosen to demonstrate the effect, but this value is compatible with the recommended smallest number of observations (\( N = 10 \)) for extreme value analysis of annual maxima (Lieblein 1974). In developed countries, observational periods are typically 20–50 years, but many locations in developing countries have observational periods that are too short for reliable analysis.

8. Conclusions

There are two fundamental axioms that stem directly from the conventional frequency interpretation of probability:

1) The best estimate of any parameter in a finite-length random sample is always the expectation of that parameter.
2) Sampling methods, or bootstrapping, exactly replicate the frequency interpretation of probability.

From axiom 1, it follows that the expectations of nonlinearly related parameters are not concomitant in any finite-length sample, which means that the best estimate of each must be separately estimated and not derived from another parameter—this is an inconvenient truth that cannot be avoided by appealing to a fundamentalist dogma. From axiom 2, it follows that mean values obtained by bootstrapping represent the expectation to any desired accuracy, subject only to the number of trials used. Bootstrapping can therefore be used to demonstrate almost any random-based issue. In addition to providing unbiased plotting positions, a valuable role of sampling methods is to provide confidence limits for any of the observations and for any value predicted from the fits.

The case made by the note appears to stem from an incomplete appreciation of the principles of extreme value analysis and of modern reliability-based design codes. The note contains selective and inaccurate reinterpretations of old sources. The fundamental error in Eq. (9) of the note erroneously selects the Weibull estimator for return period as well as for probability \( E(P) \), and this is interpreted as imbuing “uniqueness” to this estimator. So, starting with an inappropriate emphasis on return period, followed by a fundamental error in deriving its estimator, the note reduces to nine assertions—every one of which has been shown to be wrong.

Table 4 of the note wrongly set the Weibull estimator as the datum for zero error, and, as a consequence, the Harris (1996) numerical method was perversely assigned the greatest error. The opposite has been demonstrated to be true: in exactly representing the frequency interpretation of probability, sampling methods have zero bias error \( \hat{\alpha} \), and the associated variance error reduces with the number of trials \( n \): \( \alpha(\epsilon) \approx 1/n^{1/2} \). Calibration of Harris's numerical method against a \( 10^6 \)-trial bootstrap shows it to have zero bias error (within the 0.1% tolerance). Table 4 of this paper corrects the corresponding table in the note.
to this true datum, and the methods have been reordered chronologically. It is now clearly seen that successive generations of statisticians have devised estimators for the FT1 distribution that have, over time, progressively reduced bias errors to virtually zero.

Acknowledgments. The many helpful discussions with Mr. R. I. Harris of the issues addressed by this comment are gratefully acknowledged.

REFERENCES


to this true datum, and the methods have been reordered chronologically. It is now clearly seen that successive generations of statisticians have devised estimators for the FT1 distribution that have, over time, progressively reduced bias errors to virtually zero.

Acknowledgments. The many helpful discussions with Mr. R. I. Harris of the issues addressed by this comment are gratefully acknowledged.

REFERENCES


