Leadership in Sustainable Infrastructure

Leadership en Infrastructures Durables

Vancouver, Canada

May 31 – June 3, 2017/ *Mai 31 – Juin 3, 2017*

FREQUENCY ANALYSIS AND PLOTTING POSITIONS

Bertrand Massé [1, 2]
[1] SNC-Lavalin, Canada
[2] bertrand.masse@snclavalin.com

ABSTRACT

Various formulas have been proposed for estimating the probabilities of occurrence or return periods of the various elements comprised in a sample of historic hydrological data. Most of these formulas are based on the equation P = (m-α)/(N+1-2α), in which *m* is the rank of the element; *N* is the number of elements in the ranked sample; and *α* is a parameter whose value has been a subject of dispute among hydrologists for several decades. Proposed values of *α* vary between 0 and 0.5 and, presumably, depend on the probability distribution considered. A visual representation of the return period of the most extreme value in a ranked series is proposed, which shows that the most appropriate value of the parameter *α* is *α* = 0.31 and that it's independent of the probability distribution.

RÉSUMÉ

Différentes formules ont été proposées pour estimer les probabilités empiriques ou périodes de récurrence pour des échantillons de variables hydrologiques. La plupart de ces formules prennent la forme P = (m-α)/(N+1-2α), dans laquelle m est le rang d'une observation donnée, N est le nombre d'éléments dans l'échantillon et *α* est un paramètre dont la valeur a suscité beaucoup de discussions au cours des années. Les valeurs proposées de *α* varient entre 0 et 0.5 et dépendent présumément de la distribution statistique considérée. Une représentation visuelle de la période de récurrence de l'événement le plus grand dans un échantillon ordonné est présentée, laquelle montre que la valeur la plus appropriée du paramètre *α* est *α* = 0.31 et que cette valeur est indépendante de la distribution considérée.

Keywords: - Frequency analysis, plotting positions, probability, return period

1. INTRODUCTION

Frequency analysis is commonly used to estimate return periods of meteorological or hydrological variables such as precipitation or river discharge. Nowadays, the analysis is carried out using specialized software such as HYFRAN-PLUS (Bobée and El Adlouni 2015), PeakFQ (Veilleux *et al.* 2014) or HEC-SSP (USACE 2016). The software often proposes several statistical distributions which can be tested and the user has to select the distribution that offers the best fit to the actual data. The selection is made by comparing the various distributions using a special probability paper. Results of an analysis using the normal probability distribution will plot as a straight line on a normal probability paper. A straight line will also be obtained with the lognormal distribution when the variables are logarithmically transformed. Similarly, the results of the application of the Gumbel distribution will also plot as a straight line on a Gumbel probability paper. All other distributions will plot as curved lines, whether concave upward or concave downward.

Two-parameter distributions like those mentioned above most of the time do not fit well hydrological data and this explains why three-parameter distributions (generalized extreme value (GEV), log-Pearson III, etc.) were introduced. When plotted on normal or Gumbel probability paper these distributions most of the time show curved lines.

2. PLOTTING DATA ON A FREQUENCY PLOT

Let a sample of N values of a variable X be ranked in order of magnitude from the largest to the smallest:

$x_1 \geq . . . \geq x_m \geq. . . \geq x_N$

If $y_m = F(x_m)$ is the cumulative distribution function of X then the probability of exceedence of variable x_m will be decreasing:

$y_1 \geq . . . \geq y_m \geq. . . \geq y_N$

Fitting data to a frequency plot requires that a probability be assigned to each element of the sample. Gumbel (1958) has proposed the following conditions which should be satisfied for any plotting formula:

1. The plotting positions must be such that all members can be plotted;
2. The return period (plotting position) of a value equal to or larger than the largest observation should converge towards N, the number of observations;
3. The observations should be equally spaced on the frequency scale;
4. The plotting position formula ought to be simple and have an intuitive meaning; and
5. The plotting positions of observation i should lie between the observed frequencies (m-1)/N and m/N and should be universally applicable, i.e. it should be distribution-free.

Gumbel has not offered any proofs as to the necessity of the conditions. Cunnane (1978) has discussed them and expressed disagreements about conditions (2), (3) and (4). For example, he states that condition (2) is not in keeping with statistical fact and is misleading. He also questioned the necessity and desirability of conditions (3) and (4) but did not reject them.

A general plotting formula for ranked observations is the following:

[1] $$P = \frac{m-\alpha}{N+\beta}$$

in which P is the probability of occurrence of observation m; m is the rank of the observation; N is the number of observations in the sample; and α and β are parameters to be determined. In order to estimate the value of parameter β one assumes that the probabilities will be symmetrically distributed around probability 0.5. If the observations are ranked from the largest (m = 1) to the smallest (m = N), this means that the probability of exceedence of observation m will be equal to the probability of non-exceedence of observation N - (m - 1), or

[2] $$\frac{m-\alpha}{N+\beta} = 1 - \frac{N-(m-1)-\alpha}{N+\beta}$$

Resolution of the equation yields β = 1 - 2α and the plotting formula becomes:

[3] $$P = \frac{m-\alpha}{N+1-2\alpha}$$

Different values of parameter α have been proposed and Table 1 presents some of the commonly used formulae.

Hazen’s formula, Eq. 4, is more than a century old and is not based on a theoretical development but was rather proposed to circumvent the ambiguity posed by the then familiar equations m/n and (m-1)/N. Hazen avoided this problem by using $\alpha = 0.5$.

The Weibull formula, Eq. 5, has long been used because it was said to be unbiased for any statistical distribution but Cunnane (1978) showed that it gives a biased overestimation at the upper end of a skewed distribution. In fact, the Weibull formula is unbiased only for the uniform distribution.

Blom (1958) showed that for the normal distribution, an optimal unbiased estimate of the standard deviation σ may be obtained by using $\alpha = 0.375$ (Eq. 6). The same formula would also apply to the two-parameter lognormal distribution when the variables are logarithmically transformed. Using a similar method, Gringorten (1953) proposed to use $\alpha = 0.44$ (Eq. 8) for the exponential (or extreme value) distribution. Cunnane (1978) suggested using $\alpha = 0.40$ (Eq. 10) as a compromise if the formula is to be used with all distributions.

Tukey (1962) thought that the differences between the various equations is probably not important and proposed Eq. 7 because it is simple and represents an adequate approximation to what is claimed to be an optimum.

A computation method equivalent to Eq. 9 was initially proposed by Beard (1943), based on the idea that a natural estimate of the unknown variable $y_m = F(x_m)$ is the median of its distribution, the value with equal probabilities of being exceeded or not being exceeded. Later on, Chegodaev (1953) and Benard and Bos-Levenbach (1953) both used a variant of Eq. 9 where α has been rounded up to 0.30 instead of 0.31. Eq. 9 has been adopted by Jenkinson (1977) based on the same principle.

Table 1. Plotting position formulae

Equation	Plotting position	Formula	Value of α
[4]	Hazen (1914)	$P = \frac{m-0.5}{N}$	0.5
[5]	Weibull (1939)	$P = \frac{m}{N+1}$	0
[6]	Blom (1958)	$P = \frac{m-0.375}{N+0.25}$	0.375
[7]	Tukey (1962)	$P = \frac{m-0.33}{N+0.33}$	0.33
[8]	Gringorten (1963)	$P = \frac{m-0.44}{N+0.12}$	0.44
[9]	Jenkinson (1977)	$P = \frac{m-0.31}{N+0.38}$	0.31
[10]	Cunnane (1978)	$P = \frac{m-0.4}{N+0.2}$	0.4

More recently, In-Na and Nguyen (1989) have modified the general plotting formula to include the coefficient of skewness while investigating three-parameter distributions. For the generalized extreme value (GEV) distribution they established the following approximately unbiased plotting formula:

$$[11] \quad P = \frac{m - 0.13\tau - 0.27}{N - 0.08\tau + 0.38}$$

in which τ is the coefficient of skewness.

At the same time, Nguyen *et al.* (1989) developed a similar unbiased formula for the Pearson III distribution:

$$[12] \quad P = \frac{m - 0.42}{N + 0.3\tau + 0.05}$$

Eq. 12 would also apply to the logPearson III distribution when the variables are logarithmically transformed.

3. VISUAL INTERPRETATION OF THE PLOTTING FORMULAE

Only the plotting formulae conforming to the general formula expressed by Eq. 3 are considered in the following sections.

The different plotting formulae can be visualized as follows:

Let us assume a hydrological event, say peak instantaneous flow, which has a return period equal to T_r. Then the probability that this event will be exceeded on any year is $1/T_r$ and the probability it will not be exceeded is $1 - 1/T_r$. Now if we have a sample of N annual observations of that event, the probability that all these events will have a return period less than T_r will be $(1 - 1/T_r)^N$. Therefore, the probability that at least one observation in that sample will have a return period exceeding T_r will be:

$$[13] \quad P = 1 - (1 - 1/T_r)^N$$

The latter expression has been computed for different values of the return period, T_r, and of the sample size, N. The results are shown in Table 2 and illustrated on Figure 1.

Table 2. Probability that a flood of return period T_r years will be exceeded at least once over a period of N years

Return period (years)	Sample size, N					
	10	20	50	100	1000	10000
2	0.999	1.000	1.000	1.000	1.000	1.000
5	0.893	0.988	1.000	1.000	1.000	1.000
10	0.651	0.878	0.995	1.000	1.000	1.000
20	0.401	0.642	0.923	0.994	1.000	1.000
50	0.183	0.332	0.636	0.867	1.000	1.000
100	0.096	0.182	0.395	0.634	1.000	1.000
200	0.049	0.095	0.222	0.394	0.993	1.000
500	0.020	0.039	0.095	0.181	0.865	1.000
1000	0.010	0.020	0.049	0.095	0.632	1.000
2000	0.005	0.010	0.025	0.049	0.394	0.993
5000	0.002	0.004	0.010	0.020	0.181	0.865
10000	0.001	0.002	0.005	0.010	0.095	0.632

3.1 Relationship between the return period of the largest element and the sample size

The relationship between the return period of the largest event (m = 1) in a sample of N elements for the different plotting formulae is obtained as follows:

[14] $$\frac{1-\alpha}{N+1-2\alpha}=\frac{1}{T_r}$$

which can be rearranged as:

[15] $$\frac{T_r}{N}=\frac{1-1/N-2\alpha/N}{1-\alpha}$$

For large samples, Equation 15 yields

[16] $$\frac{T_r}{N}=\frac{1}{1-\alpha}$$

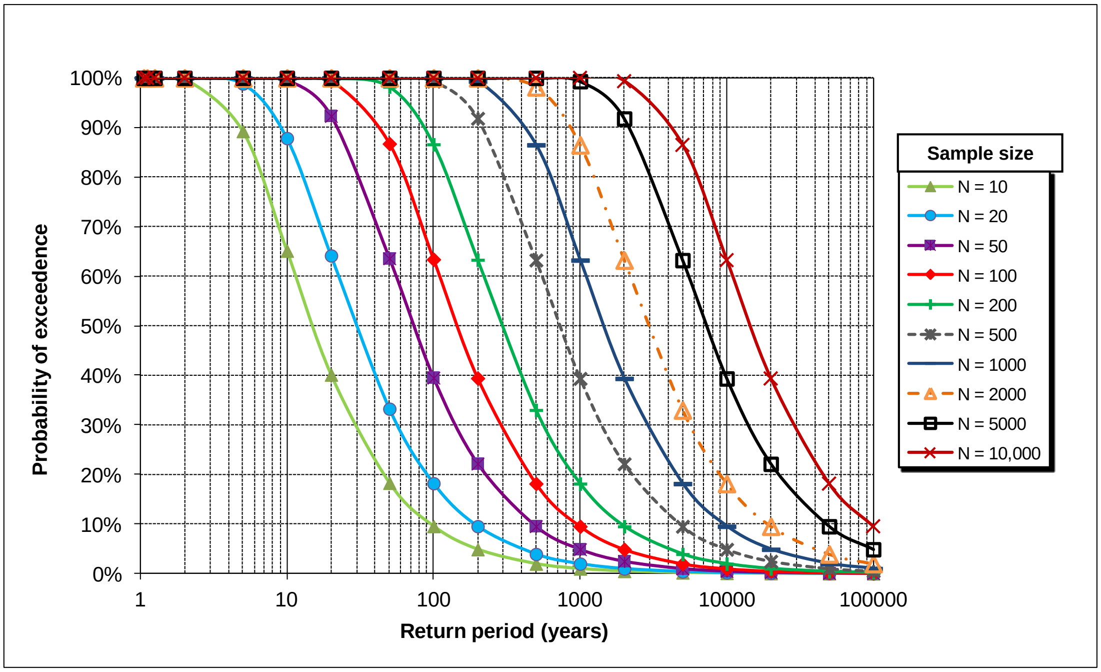

Figure 1. Probability that the T_r return period event will be exceeded at least once over a period of N years

The ratios T_r/N for the largest element in the sample as a function of the plotting formulae are shown in Table 3.

Table 3. Ratios Tr/N and probabilities of exceedence of the largest element in large samples

Formula	Tr/N	Probability of exceedence of largest element
Weibull	1.00	0.632
Jenkinson	1.44	0.500

Tukey	1.50	0.487
Blom	1.60	0.465
Cunnane	1.67	0.451
Gringorten	1.79	0.428
Hazen	2.00	0.393

3.2 Probability of exceedence of the largest element in a sample

The probability of exceedence of the largest element in the sample is given by Eq. 13. When combined with Eq. 15 this probability can be estimated for different sample sizes and the results for the various plotting formulae are illustrated in Figure 2.

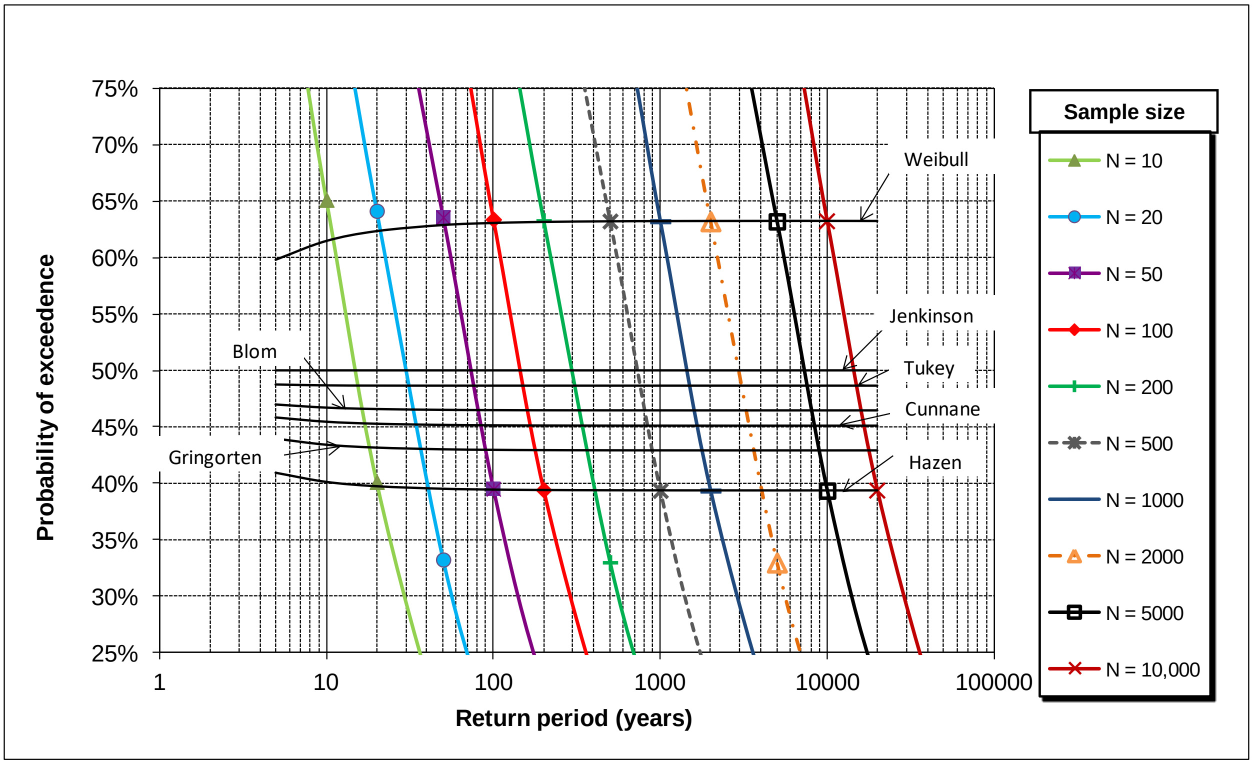

Figure 2. Probability of exceedence of the largest element for the various plotting formulae

Figure 2 shows clearly that the probability of exceedence of the largest event using Weibull's formula is over 60% which means that this formula tends to overestimate the magnitude of the event for a given return period. On the contrary, all other formulae, but Jenkinson's, tend to more or less underestimate the event. Hazen's formula is the most extreme in that regard. Only Jenkinson's formula appears neutral.

Even though Blom's or Gringorten's formula have been declared optimal for the normal or extreme value distributions, respectively, a hydrologist never knows which distribution will provide the best fit for the data and, therefore, there is no statistical reason to adopt these formulae a priori. Jenkinson's formula, by definition, can be applied whatever the statistical distribution of the sample.

It is possible that a formula other than Jenkinson's may provide a better fit for a given sample but, when used with a large number of samples, Jenkinson's formula is the one that will provide, most of the time, values for a given return period which are neither overestimated nor underestimated.

Combining Equation 16 with Equation 13 for large samples is equivalent to

$$[17]\quad P = 1 - e^{-N/T_r} = 1 - e^{\alpha - 1}$$

Probabilities of exceedence of the largest element in the sample and for the various plotting formulae are given in Table 3.

Figure 3 shows a comparison of Jenkinson's formula with other ranking formulae for a sample of size N = 50 using return periods for $46 \le m \le 50$. It is clear that a significant difference for return periods is only apparent for the two or three highest ranks. Weibull's formula gives clearly a lower return period for the largest event. All other formulae give relatively similar results except Hazen's which gives the highest estimate.

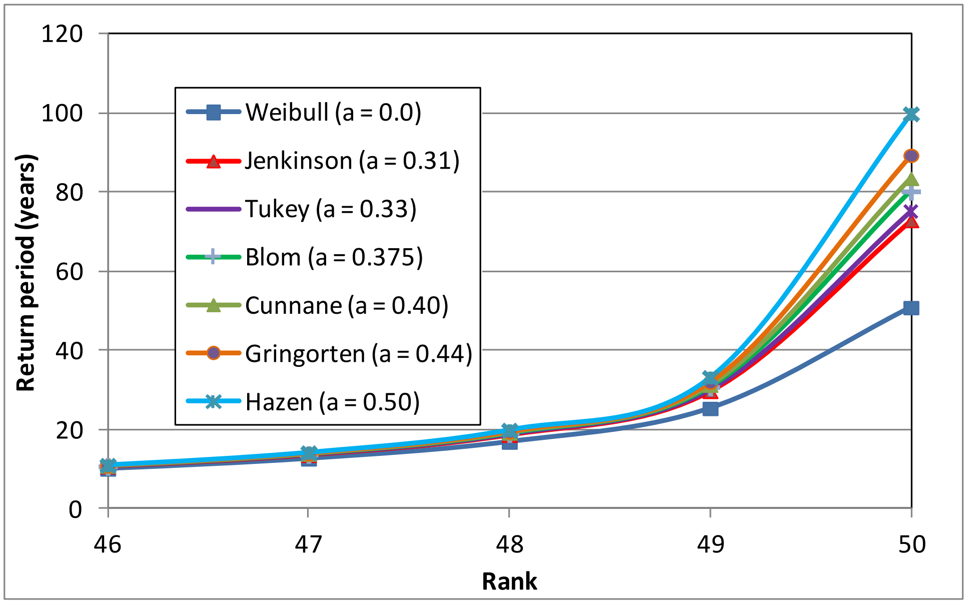

Figure 3. Comparison of Jenkinson's formula with other ranking formulae

4. THE JENKINSON FORMULA

The Jenkinson formula is different from the other plotting formulae because it is the only one that is defined based on a specific probability which corresponds to the median. In other words, the probability of an element obtained with the Jenkinson has 50% chance of being exceeded and 50% of not being exceeded. It is, therefore, independent of the statistical distribution.

Combining Eq. 13 and Eq. 14 and setting P = 0.5, one gets

$$[18]\quad \alpha = \frac{N\left(2^{1/N} - 1\right) - 1}{2^{1/N} - 2}$$

Table 4 shows the value of α for different sample sizes.

Table 4. Values of α for different sample sizes

Sample size	α
10	0.3041
50	0.3063
100	0.3066
1000	0.3068

It can be seen that for small samples ($N < 20$) α is about equal to 0.30, the value proposed by Chegodaev (1953) and Benard and Bos-Levenbach (1953). For $N \geq 20$ the value of α is closer to 0.31.

Using Equation 17 with $P = 0.5$ it can be shown that for large N, $T_r/N = 1 / \ln 2$. Similarly, Eq. 17 can be used to get an estimate of parameter α for large sample sizes: $\alpha = 1 - \ln 2$, a result already obtained by Folland and Anderson (2002).

5. CONCLUSION

Jenkinson's formula is the only one having a statistical meaning and being applicable with any statistical distribution. All other formulae generally either overestimate or underestimate the magnitude of the largest element in a sample. It is still possible that a formula other than Jenkinson's may provide a better fit for a given sample but when performing frequency analyses for a large number of samples the Jenkinson formula is likely to provide, most of the time, values for a given return period which are neither overestimated nor underestimated.

6. REFERENCES

Beard, L.R., 1943. Statistical analysis in hydrology. *Trans. Am. Soc. Civ. Eng.*, 108 : 1110-1160.

Benard, A. and Bos-Levenbach, E.C., 1953. The plotting of observations on probability paper (in Dutch). *Statistica Neerlandica*, 7: 163-173.

Blom, G., 1958. *Statistical Estimates and Transformed Beta Variables*. Wiley, New York, N.Y., pp.68-75 and 143-146.

Bobée, B. And El Adlouni, S., 2015. *Éléments théoriques d'analyse fréquentielle - Utilisation du logiciel HYFRAN-PLUS*. INRS-ÉTÉ, Québec.

Chegodaev, N.N., 1953. *Computation of surface runoff on small catchments* (in Russian). All-Union Scientific Research Institute for Railway Construction and Planning. Rep. 37, State Rail Transport Publishing House, 76 p.

Cunnane, C., 1978. Unbiased plotting positions – a review. *Journal of Hydrology*, 37: 205-222.

Folland, C. And Anderson, C., 2002. Estimating Changing Extremes Using Empirical Ranking Methods. *Journal of Climate.* (15): 2954-2960.

Gringorten, I.I., 1963. A plotting rule for extreme probability paper. *J. Geophys. Union*, 68(3): 813-814.

Gumbel, E.J., 1958. *Statistics of Extremes*. Columbia University Press, New York, N.Y.

Hazen, A., 1914. Storage to be provided in impounding reservoirs for municipal water supply. *Trans. Am. Soc. Civ. Eng*., Paper 1308, 77: 1547-1550.

In-Na, N. And Nguyen, V.T.V., 1989. An Unbiased plotting position formula for the general Extreme Value distribution. Journal of Hydrology, 106 : 193-209.

Jenkinson, A.F., 1977. The analysis of meteorological and other geophysical extremes. *Met. Office Synoptic Climatological Branch Memo*. 58, 41 pp.

Nguyen, V.T.V., In-Na, N. and Bobée, B., 1989. New plotting position formula for Pearson type-III distribution, J. Hydraul. Eng., 115 : 709-730.

Tukey, J.W., 1962. The future of data analysis. *Ann. Math. Stat*., 33(1):21-24.

US Army Corps of Engineers, 2016. *HEC-SSP Statistical Software Package V. 2.1*. *User's manual*. Institute for Water Resources, Hydrologic Engineering Center, CPD-86, Davis, CA.

Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason, R.R., Jr., and Hummel, P.R., 2014, *Estimating magnitude and frequency of floods using the PeakFQ 7.0 program*. U.S. Geological Survey Fact Sheet 2013-3108, 2 p., http://pubs.usgs.gov/fs/2013/3108/

Weibull, W., 1939. A statistical theory of strength of materials. *Ing. Vetenskaps Akad.*, Handlingar NR 51, Stockolm.