

A Note on the Plotting Position Controversy and a New Distribution-free Formula

Pasquale Erto¹ and Antonio Lepore²

Abstract The recent work “Bringing closure to the plotting position controversy” (Makkonen, 2008a) states the impossibility of a different solution from the classical plotting position proposed by Weibull (1939) and promoted by Gumbel (1958). In this paper, a new and better distribution-free solution is obtained via axiomatic approach. The applicative implications of a comparative study, carried out by using a wide Monte Carlo simulation, are presented. The new solution advantages are confirmed in estimating the median Return Period.

Keywords: Graphics and Data Visualization, distribution-free, return period.

1 Introduction

Any item must generally ensure a specified reliability level even before it is completely realized. Also for this reason, the estimation of the return period of an extreme event (*i.e.*, the inverse of the probability of such event) is a basic and a strategic task in many applicative fields, e.g., engineering and environmental sciences. In such critical contexts, practitioners find the graphical estimation process more friendly and under control, especially when the data are scarce.

¹Pasquale Erto

Department of Aerospace Engineering, University of Naples Federico II
P.le Tecchio 80, 80125 Naples, Italy
email: ertopa@unina.it

²Antonio Lepore

Department of Aerospace Engineering, University of Naples Federico II
P.le Tecchio 80, 80125 Naples, Italy
email: antonio.lepore@unina.it

The visualization of the data is generally performed by using probability papers that allow a graphical analysis of the fit provided by the selected model and an understanding of visual information otherwise easily wasted. The basic problem of the graphical estimation is how to determine the best estimate $\hat{F}_i$ of the cumulative distribution function (cuff) (i.e., the best plotting position) corresponding to the i -th order statistic $X_{(i)}$. In particular, the appropriate plotting position choice becomes crucial when estimating the return period (i.e., inverse of the probability of exceedance a fixed value x_r). If the parent distribution is known it is trivial to observe that, exact unbiased plotting positions can be obtained via order statistics-theory. Unfortunately, they are usually complex to be determined, thus invalidating the practical aim of the graphical estimation.

However, the practical purposes of the plotting position choice often overlook theoretical point of view. Engineers, hydrologists and statisticians have discussed probability plotting positions for many years. Harter (1984) provided a detailed review of the main plotting positions up to 1984, concluding that the optimum choice of plotting position depends on the distribution of the variable under consideration (i.e., the parent distribution) and on the use that has to be made of the results. It is worth mentioning that extremes are generally estimated by utilizing the theory developed by Fisher and Tippett (1928). The Generalized Extreme Value (GEV) distribution is often adopted since it is linked to the three asymptotic distributions of extremes (Fréchet, Weibull and Gumbel distributions) by a shape parameter (Kotz and Nadarajah, 2000). In (Erto, 1977) the possibility of obtaining linear estimates of Gumbel distributions parameters, whose precision is comparable with the Maximum Likelihood ones, was outlined.

On the other hand, thank to Makkonen (2008a) who endeavours to bring closure to the plotting position controversy, the question of determining distribution-free plotting positions has been lighted again. Despite of the above approach, Makkonen (2008a, 2008b) states that the probability of non-exceedance of the i -th value in N order ranked events equals

$$\hat{F}_i = \frac{i}{N+1} \quad i = 1, \dots, N \quad (1)$$

the so-called Weibull formula. The same author concludes that, regardless of the parent distribution and the application, the many other suggested plotting formulas and numerical methods to determine them should thus be abandoned.

The crux is that ideally the parent distribution does not necessarily need to be known (Yu, Huang, (2001)); in fact, probability papers are often utilized to determine the underlying distribution.

2 New distribution-free Plotting Position

In order to be useful, the plotting position must have the following properties (partly already postulated by Gumbel)

- 1) The plotting position should be analytically simple and should avoid the use of particular tables.
- 2) The plotting position should be adequate to also represent censored samples.
- 3) The plotting position should ensure that the whole available sample could be reported on the graph.
- 4) The estimate of $F(X_{(i)})$, corresponding to the median observation, should equal exactly 0.5.
- 5) The return period $T_N = 1/[1 - F(X_{(N)})]$ of a value larger than the largest observation $X_{(N)}$ should exactly equal its median value $\tilde{T}_N$ (which is independent from the parent distribution) (Lepore, 2009)

The practical form

$$\hat{F}_i = \frac{i - A}{N + 1 - 2A} \quad i = 1, \dots, N; \quad A \geq 0 \quad (2)$$

already proposed by Blom (1958) satisfies the first four properties listed previously. Moreover imposing that formula (2) satisfies also property 5, the following value for the constants A is deduced (Lepore, 2009)

$$A = N + \frac{N-1}{2^{1/N} - 2} \quad (3)$$

which results in the new plotting position

$$\hat{F}_i = \frac{N - i - 2^{-1/N}(1 - 2i + N)}{N - 1} \quad i = 1, \dots, N \quad (4)$$

3 Calculation and Analysis of the Results

For comparison purpose, the plotting position ability in estimating the median value $\tilde{T}_i$ of the return period

$$T_i = 1/[1 - F(X_{(i)})] \quad i = 1, \dots, N \quad (5)$$

of a value larger than the i -th order statistic is tested. The adopted criteria to test plotting positions is the value of the Mean Absolute Relative Error (*MARE*) of the median return period.

$$MARE = \frac{1}{N} \sum_{i=1}^N \left| \frac{\tilde{T}_i - \frac{1}{1 - \hat{F}_i}}{\tilde{T}_i} \right| \quad (6)$$

In order to investigate the predicted performances of the graphical estimates based on the proposed plotting position, 10000 pseudo-random samples having dimension $N = 5, 15, 30, 50$ are generated and ordered from three typical underlying distributions (Weibull, Gumbel and Normal).

$\tilde{T}_i$ is calculated as the median value of return period of the i -th pseudo-random ordered variable, obtained from the 10000 pseudo-random samples grouped by position. Furthermore, in Figure 1 and Figure 2 box plots of the return period (for $N = 5, 15$) are drawn in order to visualize simultaneously the plotting positions behaviour in estimating the return period as well. The value of the $MARE$'s corresponding to the plotting position (1) and the one proposed in equation (4) are compared. The less is the $MARE$, the more the plotting position is to be preferred.

It is worth pointing out the need to refer to $\tilde{T}_i$, rather than to the expected value $E[T_i]$. In fact, for a random sample $X_1, \dots, X_N$ from a continuous population with cdf $F(x)$ and probability density function (pdf) $f(x)$, it results

$$\begin{aligned} E[T_i] &= i \binom{N}{i} \int_{-\infty}^{\infty} [F(x)]^{i-1} [1-F(x)]^{N-i-1} f(x) dx \\ &= i \binom{N}{i} \int_0^1 [F]^{i-1} [1-F]^{N-i-1} dF \end{aligned}$$

By using the Beta function (e.g., Davis, 1972), it follows

$$E[T_i] = i \binom{N}{i} \frac{\Gamma(i) \Gamma(N-i)}{\Gamma(N)} = \frac{N}{N-i} \quad (7)$$

that is not finite if $i = N$.

3.1 Weibull Distribution

The values of $MARE$ in the case of a Weibull distribution:

$$F(x) = 1 - \exp\left[-(x/\alpha)^\beta\right] \quad x > 0; \alpha, \beta > 0 \quad (8)$$

with parameters $\alpha = 1$ (scale) and $\beta = 2$ (shape) are calculated and reported in Table 1.

3.2 Gumbel Distribution

The values of $MARE$ in the case of a Gumbel distribution:

$$F(x) = \exp\{-\exp[-(x-a)/b]\} \quad 0 < a, x < +\infty; b > 0 \quad (9)$$

with parameters $a = 100$ (location) and $b = 20$ (scale) are calculated and reported in Table 1.

3.3 Normal Distribution

The values of $MARE$ in the case of a Normal distribution with parameters $\mu = 100$ (mean) and $\sigma = 25$ (standard deviation) are calculated and reported in Table 1.

Table 1: $MARE$ for the Weibull, Gumbel and Normal distribution

			plotting position	
MARE of the median return period			Erto-Lepore	Weibull
distribution	Weibull ($\alpha = 1$, $\beta = 2$)	$N = 5$	0.002	0.069
		$N = 15$	0.004	0.041
		$N = 30$	0.003	0.027
		$N = 50$	0.003	0.019
	Gumbel ($a = 100$, $b = 20$)	$N = 5$	0.005	0.068
		$N = 15$	0.004	0.043
		$N = 30$	0.003	0.029
		$N = 50$	0.004	0.022
	Normal ($\mu = 100$, $\sigma = 25$)	$N = 5$	0.005	0.068
		$N = 15$	0.002	0.043
		$N = 30$	0.003	0.028
		$N = 50$	0.002	0.019

4 Conclusion

In this paper, a new distribution-free plotting position is proposed and compared with the Weibull one, marked by Makkonen (2008a, 2008b) as the best and appropriate plotting position. Even so, a wide Monte Carlo simulation shows that the proposed plotting position performs always better than the Weibull ones in estimating sample return periods. In particular, the best results are found when samples are very small ($N = 5, 15$) both in the case of skewed models (such as Weibull and Gumbel distribution) and in the case of a symmetrical one (Normal distribution).

Therefore, to graphically estimate the Return Period, the new distribution-free plotting position is generally suggested in all those situations where data are very scarce.

Figure 1: Box plot of the Return Period T_i $i = 1, \dots, 5$

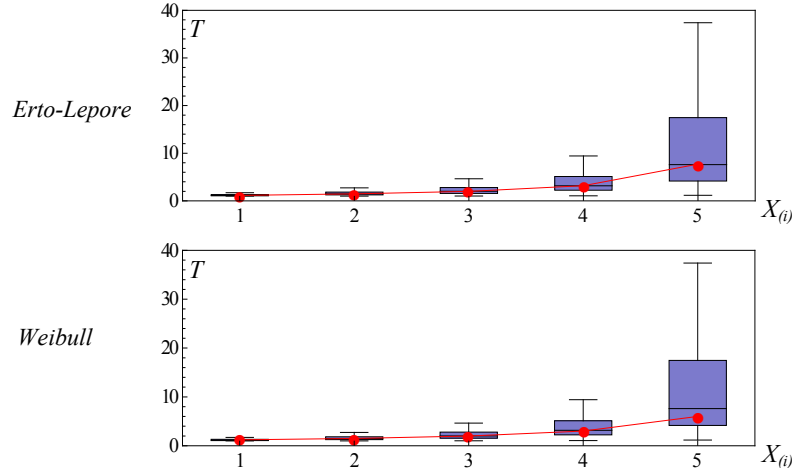
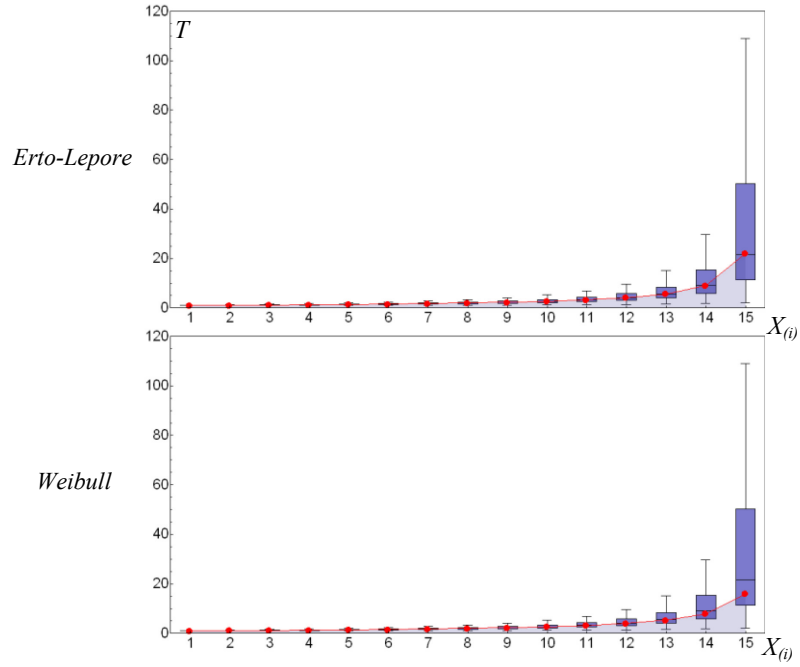


Figure 2: Box plot of the Return Period T_i $i = 1, \dots, 15$



Acknowledgements

We are grateful to the referee for the useful and pertinent suggestions that helped us to improve the manuscript significantly.

References

1. Blom, G.: Statistical estimates and transformed beta variables. New York: N.Y. Wiley, (1958)
2. Erto, P.: Stimatori lineari semplici dei parametri di affidabilità della distribuzione Gumbel da campioni censurati, Riv. Statistica **37**(1), 41-49 (1977)
3. Erto, P., Lepore A.: A step ahead toward the best distribution-free plotting position (2010, *in progress*)
4. Fisher, R.A., Tippett, L.H.C.: Limiting forms of the frequency distributions of the largest or smallest members of a sample. Math. Proc. Camb. Philos. Soc. **24**:180–190 (1928)
5. Gumbel, EJ: Statistics of extremes. Columbia Univ Press (1958)
6. Kotz, S., Nadarajah, S.: Extreme Value Distributions, Theory and Applications. London: UK, Imperial College Press, 61-95 (2000)
7. Lepore, A.: An integrated approach to shorten wind potential assessment, PhD Thesis, University of Naples Federico II, Department of Aerospace Engineering, Naples (2010)
8. Makkonen, L.: Plotting Positions in Extreme Value Analysis. J. Appl. Meteorol. Climatol. **45**, 334–340 (2006)
9. Makkonen, L.: Bringing Closure to the Plotting Position Controversy. Commun. Stat.-Theory Methods **37**, 460–467 (2008a).
10. Makkonen, L.: Problems in the extreme value analysis. Struct. Saf. **30**(5), 405–419 (2008b) doi: 10.1016/j.strusafe.2006.12.001
11. Weibull, W.: A statistical theory of strength of materials. Ing Vet Ak Handl (Stockholm) 151 45 (1939)
12. Yu, G.-H., Huang, C.-C.: A distribution free plotting position. Stoch. Environ. Res. Risk Assess. **15**(6), 462-476 (2001) doi:10.1007/s004770100083
13. Davis, P.J.: Gamma Function and Related Functions in: Abramowitz, M., Stegun, I.A. (eds.) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. p. 258. Dover, New York (1972)