Structural Safety 51 (2014) 1–12

Contents lists available at ScienceDirect

Structural Safety

journal homepage: www.elsevier.com/locate/strusafe

On the physical and probabilistic consistency of some engineering random models

Enrique Castillo [a,*], Alan J. O'Connor [b], María Nogal [b], Aida Calviño [a]

[a] Department of Applied Mathematics and Computational Sciences, University of Cantabria, 39005 Santander, Spain
[b] Department of Civil Engineering, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Article history:
Received 24 March 2014
Received in revised form 12 May 2014
Accepted 15 May 2014
Available online 24 June 2014

Keywords:
Statistical models
Operation stability
Location-scale families
Extreme value distributions
Conditional specification of joint distributions

ABSTRACT

In this paper we deal with the probability and physical consistency of random variables and models used in engineering design. We analyze and discuss the conditions for a model to be consistent from two different points of view: probabilistic and physical (dimensional analysis). The first leads us to the concept of probabilistically consistent models, which arises when the joint distributions of all variables are required. This implies that relations among the variables must be respected by densities and resulting moments. In particular the most common linear, product and quotient relations, which are physically justified, must be especially considered. Similarly, stability with respect to minimum or maximum operations and consistency with respect to extremes (maxima and minima) arises in practice. From the dimensional analysis point of view, some models are demonstrated to be inconsistent. In particular, log-normal and chi-squared models are shown to be non adequate for location or location-scale variables. The problem of building compatible models based on conditional distributions and regression functions is analyzed too. It is shown that incompatible models can be easily obtained if a consistency analysis is not performed. All these and other problems are discussed and some models in the literature are analyzed from these two points of view. When some families fail to satisfy the desired properties, alternative models are provided. Finally, some simple examples and conclusions are given to summarize the analysis.

1. Introduction and motivation

When working with statistical models in engineering and other areas we have to face different ways of solving the problems and different assumptions, data and models. Though most of them are satisfactory, we also encounter cases in which important mistakes are made. The main cause for this situation is that modeling in Engineering is a difficult and interdisciplinary task and requires knowledge from many and very different areas. This collection of errors and mistakes, from which nobody is free of, has motivated this paper. Our main aim is to point out some of the errors encountered in practice and give orientations on how to avoid them. This is not an easy task, but we want to contribute with our sand grain and to motivate other researchers to join our adventure. To clarify what we have in mind, we start by including a short list of examples below.

Example 1.1 (*Inadequate formula*). Our first example deals with the validity of some formulas. Some authors suggest to evaluate the gas emissions produced in a road by the formula:

$$E = C_1 + C_2 a v^2, \tag{1}$$

where $C_1 = 0.0039$, $C_2 = 0.0088$ are constants, E is the emission (in g/s), and a and v are the instantaneous acceleration and speed, respectively. Formula (1) presents problems:

1. A dimensional analysis, as indicated in Table 1 where L, T and M refer to the length, time and mass magnitudes, reveals that C_1 and C_2 have dimensions. Unfortunately, the authors do not indicate them and consequently the formula cannot be used. This formula is valid only for being applied to variables in given dimensions, unless the constants are recalculated.
2. The formula ignores relevant variables that must include the mass dimension. This is why the authors need to incorporate constants with dimensions.
3. Constants appearing in formulas should not be confused with variables and if possible they should be dimensionless.

Identification of the type of formulas which are valid and those which are not and how to built valid formulas is the first step in building models and deserves a careful discussion.

* Corresponding author. Tel.: +34 942201722; fax: +34 942201703.
E-mail address: castie@unican.es (E. Castillo).

http://dx.doi.org/10.1016/j.strusafe.2014.05.003
0167-4730/

Table 1
Dimensional decomposition of the variables involved in Formula (1).

Magnitude	C_1	C_2	E	a	v
L	0	−2	0	1	1
T	−1	2	−1	−2	−1
M	1	1	1	0	0

Example 1.2 (*Probability papers*). Our second example refers to the incorrect use of probability papers when they are used in extreme value analysis or to identify domains of attraction. Though it should be well known, some people seem to ignore that when dealing with extremes there are two kinds of probability papers, one for maxima (right tail) and another for minima (left tail). Using the inadequate paper can lead to serious errors and consequences. For example, when predicting minimum temperatures the use of the Gumbel probability paper plot in the upper part of Fig. 1 reveals two important errors: (a) it is a maximal probability paper when we have a problem of minima, and (b) the model is fitted to all data, when only the left tail data should be used.

Contrary, the plot in the lower part is the adequate one because it is a minimal (reverse) Gumbel probability paper. In addition, the reverse Gumbel model has been fitted using the smallest 40 data points, that is, fitting only the left tail of the data and not all the range. All the above justifies the importance of clarifying the role of probability papers and the differences when they are used (a) to identify families of parametric distributions or (b) when they are utilized for extreme value analysis.

Example 1.3 (*Temperature example*). Our third example illustrates the use of an inadequate parametric family of distributions. Consider a location with warm temperatures and suppose that the 0.98 percentile wants to be predicted. If we assume that temperatures are log-normally distributed, then we can use this model for the percentile prediction by taking a sample, calculating the logarithms and estimating their normal mean μ and standard deviation σ. Once we have these estimates, we can use the percentile formula $\log \hat{x}_{0.98} = \hat{\mu} + 2.054\hat{\sigma}$ and return to the original population by using $x_{0.98} = \exp(\log x_{0.98})$. The problem is that if a person in the US uses Farenheit temperatures the result is different from that obtained by another person, say in France, using Celsius temperatures. For example, if we use the following sample in Farenheit degrees $\{53.6, 77.0, 91.4, 62.6, 75.2, 89.6, 82.4, 57.2, 73.4, 59.0\}$, we get a percentile of 105 °F. However, using the same sample in °C, that is, {12, 25, 33, 17, 24, 32, 28, 14, 23, 15}, we get a percentile of 44.05 °C, which corresponds to 111.3 °F and not to the previous temperature 105 °F. This illustrates the inconsistency of the log-normal assumption in this particular case and the risk of predicting erroneous return periods (leading to unsafe or overly costly engineering designs). This occurs because the log-normal family of distributions is not stable with respect to changes in location. If we assume that the temperatures in Farenheit degrees are log-normal, we can obtain by a change of variable, the distribution (not a log-normal) of temperatures in Celsius degrees and even extend the log-normal to a more complicated one; however, it does not seem recommendable using a different or a complicate family for different units of measure if it is not strictly necessary. Thus, knowledge of what families of parametric distributions can and cannot be used in a given case has a high relevance in statistical modeling.

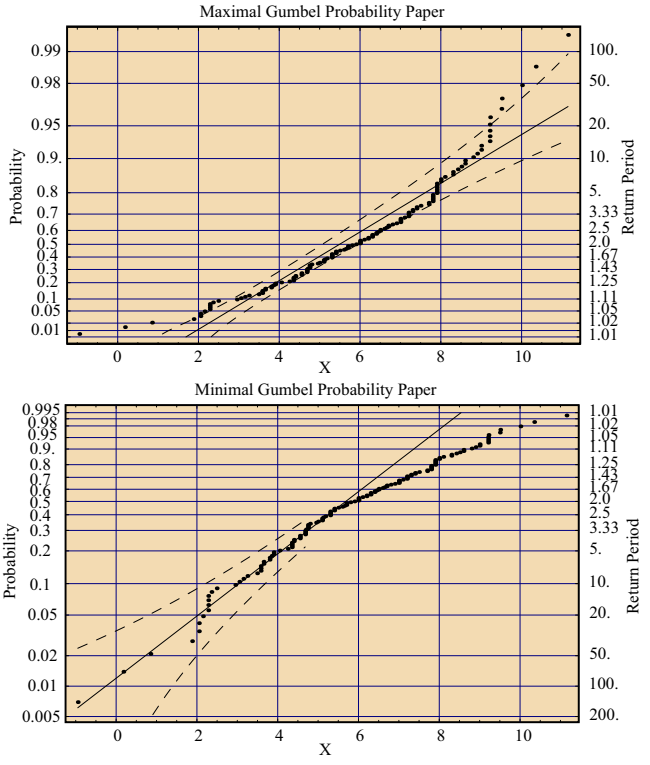

Fig. 1. Illustration of the Uppsala minimum temperatures data plotted on a maximal (upper plot) and minimal (lower plot) Gumbel probability papers.

Example 1.4 (*Wind example*). This example illustrates the use of an inadequate domain of attraction. Consider an engineering design problem in which large winds play a decisive role. Then, it is extremely important to identify the distribution of largest winds. To this end, deciding among reverse Weibull, Gumbel and Frechet is crucial. Thus, knowledge of the fact that finite or bounded variables cannot have a maximal domain of attraction of Frechet type is relevant. Selecting Frechet as the design distribution implies a very conservative position that can lead to unnecessary expenses. A more reasonable and possibly still conservative assumption is the maximal Gumbel domain of attraction (see [1] or [2]). Finally, choosing a Weibull maximal domain of attraction could be the best decision, but the corresponding parameters must be estimated with care. Thus, a good knowledge of what families of extreme value distributions can be used in a given case and why is relevant.

Example 1.5 (*Incompatible models*). Assume that we want to model the statistical behavior of a bivariate random variable (X, Y) and we decide to assume that the two conditional families $f_{X|Y}(x \mid y; \theta_x)$ and $f_{Y|X}(y \mid x; \theta_y)$ are normals and the variance of $X \mid Y$ is linear. In this case no bivariate distribution satisfies these conditions, thus, we are in front of an impossible model.

This justifies a general discussion about what conditions can and what cannot be imposed to have models compatible with the assumptions.

The previous examples illustrate the need for using consistent models. Consequently, when working with engineering stochastic models, in order to avoid serious problems it is important to check that they are probabilistically and dimensionally consistent. In this paper we deal with this topic and consider all types of models including univariate and multivariate. From the point of view of dimensional analysis, the main problem with inconsistent models is that if we use different data dimensions or different subsets of equations in the model we obtain different results. Contrary, consistent models always lead to the same results no matter what model formulas you utilize or what variable units you use in the calculations.

1.1. Aims of the paper

In this paper we analyze and discuss some probability models used to reproduce structural random variables from the point of view of probability and dimensional consistency and physical validity. The main aims of the paper are: (a) to provide examples