

Introductory statistical ideas for IOGP: distributions

Philip Jonathan and Gemini AI

September 2026

1 Introduction to univariate distributions and tails

For a continuous random variable X representing an engineering met-ocean parameter (such as wave height or wind speed), the cumulative distribution function (CDF) $F(x)$ and probability density function (PDF) $f(x)$ describe the probability characteristics of the population.

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt \quad (1)$$

The relationship between $f(x)$ and $F(x)$ is given by differentiation:

$$f(x) = \frac{d}{dx} F(x) \quad (2)$$

Domain of X : $x \in \mathbb{R}$ (or constrained to $x \geq 0$ for physical quantities like wave heights).

1.1 Tail function

The tail function $\bar{F}(x)$, also known as the survival function, represents the probability that the random variable exceeds a given value x :

$$\bar{F}(x) = 1 - F(x) = P(X > x) \quad (3)$$

Domain: x in the support of X , with $\bar{F}(x) \in [0, 1]$.

1.2 Hazard function

The hazard function $h(x)$ (or failure rate function) expresses the instantaneous probability of exceeding x given that the variable has already reached x :

$$h(x) = \frac{f(x)}{\bar{F}(x)} = -\frac{d}{dx} \ln \bar{F}(x) \quad (4)$$

Domain: $h(x) \geq 0$ for all x where $\bar{F}(x) > 0$.

1.3 Return value and annual maximum distribution

Let X_A be the annual maximum of X . Assuming the occurrences of extreme events follow a Poisson process with an expected rate $\lambda > 0$ events per annum, the distribution of X_A is derived as follows.

Analytical Demonstration: Let N be the number of events in a year, where $P(N = n) = \frac{e^{-\lambda} \lambda^n}{n!}$. Let $X_1, X_2, \dots, X_N$ be independent and identically distributed (i.i.d.) instances of X with common CDF $F(x)$. The annual maximum is $X_A = \max(X_1, \dots, X_N)$, with $X_A = -\infty$ if $N = 0$.

The CDF of X_A , denoted $G_A(x)$, is obtained by conditioning on N :

$$G_A(x) = P(X_A \leq x) = \sum_{n=0}^{\infty} P\left(\max(X_1, \dots, X_N) \leq x \mid N = n\right) P(N = n) \quad (5)$$

$$= \sum_{n=0}^{\infty} [F(x)]^n \frac{e^{-\lambda} \lambda^n}{n!} \quad (6)$$

$$= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda F(x))^n}{n!} = e^{-\lambda} e^{\lambda F(x)} \quad (7)$$

$$= \exp(-\lambda(1 - F(x))) = \exp(-\lambda \bar{F}(x)) \quad (8)$$

Thus, the distribution of the annual maximum is $G_A(x) = \exp(-\lambda \bar{F}(x))$, where parameters are $\lambda > 0$ and $\bar{F}(x) \in [0, 1]$. A T -year return value x_p corresponding to an annual exceedance probability $p = 1/T$ is defined by $G_A(x_p) = 1 - \frac{1}{T}$.

2 Bivariate distributions and copulas

For a continuous bivariate random variable (X, Y) , the joint cumulative distribution function $F(x, y)$ and joint probability density function $f(x, y)$ capture dependencies between environmental variables (e.g., wave height and storm surge).

The joint PDF can be decomposed using Sklar's theorem via marginal densities and a copula density c :

$$f(x, y) = f_X(x) f_Y(y) c(F_X(x), F_Y(y)) \quad (9)$$

where $f_X(x)$ and $f_Y(y)$ are the marginal PDFs, and $c(u, v)$ is the copula density with parameters $\theta \in \Theta$ (copula dependence parameters, e.g., $\rho \in [-1, 1]$ for Gaussian copulas).

The joint CDF is given by:

$$F(x, y) = C(F_X(x), F_Y(y)) \quad (10)$$

where $C(u, v)$ is the copula distribution function mapping $[0, 1]^2 \rightarrow [0, 1]$.

3 Conditional tail probabilities and generalised Pareto limits

The conditional tail probability for a univariate variable X exceeding a high threshold u is given by:

$$P(X > x \mid X > u) = \frac{\bar{F}(x)}{\bar{F}(u)}, \quad x > u \quad (11)$$

As the threshold $u \rightarrow \infty$, the Pickands-Balkema-De Haan theorem states that the conditional excess distribution converges to a Generalised Pareto Distribution (GPD):

$$P(X > u + y \mid X > u) \approx \left(1 + \xi \frac{y}{\tilde{\sigma}(u)}\right)^{-1/\xi} \quad (12)$$

where $y = x - u > 0$, scale parameter $\tilde{\sigma}(u) > 0$, and shape parameter $\xi \in \mathbb{R}$.

3.1 Threshold stability test

The threshold stability property implies that if X follows a GPD above threshold u_0 , then for any higher threshold $u > u_0$, the conditional excesses also follow a GPD with the same shape parameter ξ and a modified scale parameter $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$.

Algorithm to check for threshold stability:

1. Select a sequence of candidate thresholds $u_1, u_2, \dots, u_m$ at high quantiles (e.g., 80% to 98% quantiles).
2. For each threshold u_k , fit the GPD parameters $(\xi(u_k), \sigma^*(u_k) = \sigma(u_k) - \xi u_k)$ using Maximum Likelihood Estimation (MLE) on exceedances $X > u_k$.
3. Plot the modified scale parameter $\sigma^*(u_k)$ and shape parameter $\xi(u_k)$ as functions of u_k along with confidence intervals (mean residual life plot and parameter stability plots).
4. Identify the lowest threshold u^* above which $\xi(u)$ remains approximately constant and $\sigma^*(u)$ varies linearly with u .

4 Max-stability

A distribution function G is max-stable if the maximum of n independent copies, suitably normalised, belongs to the same distribution family.

$$G^n(a_n x + b_n) = G(x) \quad \text{for constants } a_n > 0, b_n \in \mathbb{R} \quad (13)$$

Domain of parameters: $a_n > 0, b_n \in \mathbb{R}$.

4.1 Equivalence of block-maxima max-stability and GPD tail

Analytical Demonstration: Let $X_1, X_2, \dots, X_n$ be i.i.d. variables with CDF F . The block maximum $M_n = \max(X_1, \dots, X_n)$ has CDF $F^n(x)$. Assuming max-stability in the limit as $n \rightarrow \infty$, $F^n(a_n x + b_n) \rightarrow G(x)$, where G is a Generalised Extreme Value (GEV) distribution:

$$G(x) = \exp \left(- \left(1 + \xi \frac{x - \mu}{\sigma} \right)^{-1/\xi} \right) \quad (14)$$

Taking the logarithm and applying a first-order Taylor expansion for large levels u , we write $1 - F(u) \sim \frac{1}{n}(-\ln G(u))$. Substituting the GEV form shows that the tail of F matches a power-type or exponential decay corresponding precisely to the tail behaviour of a Generalised Pareto distribution for exceedances $X > u$.

4.2 Generalised Pareto limit equivalent to max-stability

Step-by-step argument:

1. Assume the tail of F satisfies the GPD limit condition for exceedances over u : $\bar{F}(u + y)/\bar{F}(u) \approx (1 + \xi y/\sigma(u))^{-1/\xi}$.
2. Consider the probability that the maximum of n observations exceeds a high level x_n : $P(M_n \leq x_n) = F^n(x_n) = (1 - \bar{F}(x_n))^n$.
3. Substitute the GPD survivor function into the expression: $(1 - \frac{1}{n} \bar{F}(x_n))^n$.
4. As $n \rightarrow \infty$, using the limit $(1 - c/n)^n \rightarrow e^{-c}$, the distribution of block maxima converges to the GEV distribution family.
5. Since the limit distribution of block maxima is GEV if and only if the conditional tail converges to GPD, max-stability of block maxima is analytically equivalent to the GPD tail limit.

5 Conditional extremes model (Heffernan and Tawn)

To model multivariate extremes when one variable is extreme (e.g., surge given high wave height), Heffernan and Tawn propose a semi-parametric conditional approach.

The model requires marginal transformations of X and Y to standard Laplace scale, denoted X_L and Y_L . The conditional extremes model is expressed as:

$$X_L \mid (Y_L = y_L) = \mathbf{a}(y_L) + \mathbf{b}(y_L)Z \quad \text{for } y_L > u_L \quad (15)$$

where Z is a residual vector independent of y_L in the limit, and $\mathbf{a}(y)$ and $\mathbf{b}(y)$ are normalising functions. Standard practice assumes parametric forms:

$$a(y) = a \times y, \quad b(y) = y^b \quad (16)$$

Parameter domains: $a \in [-1, 1]$, $b \in (-\infty, 1)$, and residual Z has mean zero and finite variance (typically assumed Gaussian).

5.1 Transformation to Laplace scale

Algorithm for marginal transformation:

1. Given original observations X , compute empirical/parametric CDF $F_X(x)$.
2. Apply the Probability Integral Transform (PIT) to obtain uniform variables: $U = F_X(X) \in (0, 1)$.
3. Transform U to standard Laplace scale via the inverse Laplace CDF $F_L^{-1}(U)$:

$$X_L = \begin{cases} \ln(2U), & U \leq 0.5 \\ -\ln(2(1-U)), & U > 0.5 \end{cases} \quad (17)$$

Equation linking CDFs: $F_{X_L}(x_L) = F_L(x_L) = \frac{1}{2}e^{x_L}$ for $x_L \leq 0$, and $1 - \frac{1}{2}e^{-x_L}$ for $x_L > 0$.

5.2 Gaussian assumption for the residual process Z

For practical model fitting, the residual process Z is assumed to be normally distributed: $Z \sim \mathcal{N}(\mu_z, \sigma_z^2)$.

5.3 Algorithm for fitting procedure starting from original scales:

1. Extract raw paired data (X_i, Y_i) for $i = 1, \dots, n$.
2. Transform X and Y to uniform margins using empirical CDFs, then transform to Laplace margins (X_{Li}, Y_{Li}) using the Laplace inverse CDF.
3. Select a high threshold u_L for Y_L (e.g., the 95th percentile).
4. For all data points where $Y_{Li} > u_L$, estimate parameters a and b using weighted least squares or maximum likelihood on the regression relationship: $Z_i = \frac{X_{Li} - aY_{Li}}{Y_{Li}^b}$.
5. Verify residuals Z_i for normality (e.g., via Q-Q plots) and estimate residual mean $\hat{\mu}_z$ and standard deviation $\hat{\sigma}_z$.
6. Confirm threshold stability of parameter estimates.

6 Regular variation and tail equivalence

Regular variation provides a mathematical framework for **heavy-tailed** distributions (with $\xi > 0$). A positive random variable X has a regularly varying tail with index $\alpha > 0$ if:

$$\lim_{t \rightarrow \infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-\alpha}, \quad x > 0 \quad (18)$$

Parameter domain: $\alpha > 0$ (tail index).

Variants of regular variation can be defined for light tailed distributions $\xi \neq 0$.

6.1 Analytical equivalence of regular variation and max-stability

Demonstration: Let block maxima M_n satisfy max-stability, implying convergence to a Fréchet distribution $G(x) = \exp(-x^{-\alpha})$ for $x > 0, \alpha > 0$. By definition, $P(M_n \leq x) = F^n(x)$. For large n , $F^n(a_n x) \rightarrow G(x) \implies n\bar{F}(a_n x) \rightarrow x^{-\alpha}$. Setting a_n such that $n\bar{F}(a_n) = 1$, we obtain $\lim_{n \rightarrow \infty} \frac{\bar{F}(a_n x)}{\bar{F}(a_n)} = x^{-\alpha}$, which matches the exact definition of regular variation of $\bar{F}$ with index α .

6.2 Regular variation of the generalised Pareto tail

Demonstration: A GPD survival function for exceedances above u is $\bar{F}_u(y) = (1 + \xi y/\sigma)^{-1/\xi}$ for $\xi > 0$. Let $\xi = 1/\alpha > 0$ and hence $\alpha > 0$. Consider scaling the argument by a factor $c > 0$:

$$\lim_{u \rightarrow \infty} \frac{\bar{F}(u + cy)}{\bar{F}(u)} = \lim_{u \rightarrow \infty} \left(\frac{1 + \frac{1}{\alpha} \frac{u+cy}{\sigma}}{1 + \frac{1}{\alpha} \frac{u}{\sigma}} \right)^{-\alpha} = \left(\lim_{u \rightarrow \infty} \frac{u + cy}{u} \right)^{-\alpha} = c^{-\alpha} \quad (19)$$

This confirms that the tail of the Generalised Pareto distribution is regularly varying with index $\alpha = 1/\xi$.

7 Prompt script reproduction

The following text block reproduces the exact prompt script used to generate this document, formatted in verbatim style with word-wrapping enabled.

```
You are a university teacher of applied statistics to engineering students.

Your students have a rudimentary understanding of probability distributions and
statistical modelling.

The students need a basic understanding of statistical distributions, methods for
describing the TAILS of those distributions, and methods for fitting models to
data.

You are going to produce a LaTeX report in multiple sections.

In producing the report:
- Please confirm that all LaTeX code generated is valid, and that there are no minor
  typographical errors.
- Please provide full step by step mathematical derivations when asked to "demonstrate
  analytically" are requested.
- Please provide full step by step algorithms when asked to "Give algorithm".
- With every concept introduced, please provide a simple equation explaining the
  concept, NOT JUST WORDS.
- Please define the domains of all parameters in statistical models
```

- Please use sentence case for all titles
- Please use the report title "Introductory statistical ideas for IOGP masterclass" and sub-title "Stats for Met-ocean"
- Please state that the authors are "Philip Jonathan and Gemini AI"
- Please make sure that LaTeX is compatible with pdftex

Please provide a LaTeX report with the following sections:

1. Introduce the cumulative probability function F and the probability density function f for a continuous random variable X , and the relationships between f and F
 - 1.1 Define the tail function $\bar{F} = 1 - F$
 - 1.2 Define the hazard function $h(x) = f(x)/\bar{F}$
 - 1.3 Define a return value in terms of the distribution of the annual maximum X_A of X , where the rate of occurrence of instances of X are assumed to be Poisson-distributed with an expected number λ of events per annum
 - Demonstrate analytical how to obtain the distribution of X_A from knowledge of the distribution of X and of λ
2. Introduce the cumulative probability function F and the probability density function f for a continuous bivariate random variable (X,Y) , and the relationships between f and F
 - Provide equation for $f(X,Y)$ as the product of the marginal pdfs and a copula density
 - Provide equation for $F(X,Y)$ as the product of the marginal cdfs and a copula distribution.
3. Provide equation for the conditional tail probability $P(X > x | X > u)$
 - Provide equation for its generalised Pareto limit as $u \rightarrow \infty$
 - 3.1 Describe the threshold stability test for selection of u
 - Give algorithm to check for threshold stability
4. Describe the concept of max-stability of X
 - Give equation for max-stability
 - 4.1 Demonstrate analytically how max-stability of block-maximum of copies of X is equivalent to a generalised Pareto limit for the conditional tail of X
 - 4.2 Demonstrate analytically how a generalised Pareto limit for the conditional tail of X is equivalent to max-stability of copies of X
 - please provide a step by step argument
5. Describe the conditional extremes model of Heffernan and Tawn applied to describe $X_L | (Y_L = y_L)$ where $y_L > u_L$ for Laplace-scale transformations X_L of X and Y_L of Y
 - State that the model requires marginal transformation to Laplace scale
 - Give equation for Heffernan and Tawn model in detail
 - explain that $a(y) = a \times y$ and that $b(y) = y^b$ are usually assumed
 - give domains for parameters
 - 5.1 Describe how the transformation from original margins to Laplace-scale is achieved using 3 above and the probability integral transform.
 - Give algorithm
 - Give equation linking cdfs

- 5.2 Describes the pragmatic modelling assumption that the residual process Z is assumed Gaussian for fitting.
- Give algorithm for fitting procedure, starting with X and Y on original scales.
6. Describes regular variation of the tail of $F(X)$ in 1D and its relationship to the generalised Pareto limit
- Give equation
 - Demonstrate analytically the equivalence of the assumption of regular variation of X with max-stability of block maxima of copies of X
 - Demonstrate analytically the regular variation of the generalised Pareto tail
7. Reproduce this prompt script explicitly, so that readers can see how Gemini AI was used to generate it.
- Please use word wrapping to avoid overly long lines not formatting properly in the pdf.
 - please use VERBATIM to ensure that the prompt script is properly word-wrapped in the eventual pdf.
 - Please confirm that the .tex file compiles correctly