Data Analytics for Coastal Systems

Georgios Boumis, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
The University of Maine

© 2025. All rights reserved.
No part of this document may be reproduced without permission from the author.

Abstract. This course introduces statistical methods for analyzing environmental data, where students develop R programming skills through hands-on work on processing real-world data from Maine's coastal environments. Emphasis is given in datasets such as sea level observations, temperature and sea-surface pressure measurements, as well as seasonal wind speed records. The course advances systematically from basic data visualization and descriptive statistics to predictive modeling, probability distribution fitting via maximum likelihood estimation, and parametric uncertainty quantification. It culminates with the application of extreme-value theory and trend detection for evaluating rare coastal events, particularly within the context of non-stationarity due climate change-driven sea-level rise.

Contents

3 Common Probability Distributions, Parameters, and Moments

Statistical modeling of random variables is based primarily on the concept of **probability distributions**. These functions work by taking a numerical value as input and returning the probability (p) that our random variable equals that value. In this course, all random variables with which we will deal are **continuous random variables**, that is, they can take any value within a given interval including all possible fractional values. The daily average temperature we saw in **Chapter 2** is a good example of a continuous random variable since, in theory, it can take any real value, i.e., its sample space, denoted as Ω, is continuous.

Since for continuous random variables there are infinitely many possible values in their sample space, the probability of a continuous random variable being exactly equal to any of these values is simply zero. Therefore, it no longer makes sense to talk about probability distributions for continuous random variables, but instead we introduce here the concept of **cumulative distribution function** (cdf). The cdf (also denoted as F) helps us calculate the probability that our continuous random variable falls within an interval of its Ω, say, $[m, n]$:

$$P(m \leq X \leq n) = F(n) - F(m), \tag{1}$$

where:

$$F(x) = P(X \leq x). \tag{2}$$

F is a non-decreasing function and if the lower and upper limit of Ω is x_- and x_+, respectively, then $F(x_-) = 0$ and $F(x_+) = 1$. This practically means that there is 0% probability ($p = 0 \times 100$) of measuring a value $\leq x_-$ and 100% probability ($p = 1 \times 100$) of measuring a value $\leq x_+$. *It is important to remember that probability is a unitless quantity and always falls between 0 and 1, that is, $0 \leq p \leq 1$.*

Another useful function for continuous random variables is the so-called **probability density function** (pdf). This function takes as input a numerical value and returns the "density" of probability at that value. *It is important to remember that the units of probability density are $1/$(units of continuous random variable).* Thus, the probability density function of the daily average temperature, as an example, yields units of $1/°\text{C}$. If F is a differentiable function, then the pdf (also denoted as f) is given by:

$$f(x) = \frac{dF}{dx}, \tag{3}$$

and therefore:

$$F(x) = \int_{-\infty}^{x} f(v)dv = \int_{x_-}^{x} f(v)dv. \tag{4}$$

It follows then from Equation 4 that Equation 1 can actually be rewritten as:

$$P(m \leq X \leq n) = \int_{x_-}^{n} f(v)dv - \int_{x_-}^{m} f(v)dv, \tag{5}$$

$$P(m \leq X \leq n) = \int_{m}^{x_-} f(v)dv + \int_{x_-}^{n} f(v)dv, \tag{6}$$

$$P(m \leq X \leq n) = \int_{m}^{n} f(v)dv. \tag{7}$$

Just as we can compute sample statistics, e.g., mean and standard deviation that summarize our measurements (see **Chapter 2**), we can also summarize key characteristics of a pdf by deriving the so-called "moments"; the most common moments are: 1) **expectation**, and 2) **variance**. The expectation ($E[X]$) is

Figure 1: Close-up of meteorological/tidal station at Cutler Farris Wharf, ME, where sea-surface pressure is being measured, among other meteorological and coastal variables (© NOAA Tides and Currents).

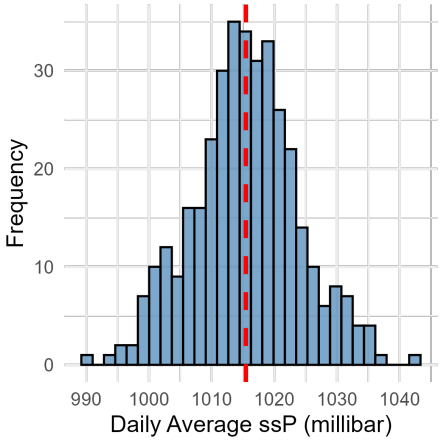

Figure 2: Histogram of daily average sea-surface pressure (ssP) measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

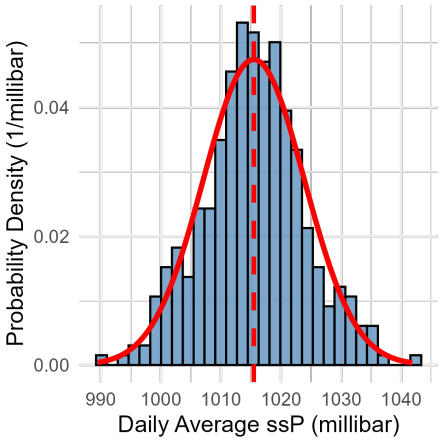

Figure 3: Probability density function of normal distribution (red solid line) fit to daily average sea-surface pressure (ssP) measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

a measure of centrality for the pdf and for a continuous random variable with probability density function f is given as follows:

$$E[X] = \int_{\Omega} xf(x)dx = \int_{x_-}^{x_+} xf(x)dx. \tag{8}$$

The variance ($Var[X]$) is a measure of spread for the pdf and for the case of a continuous random variable with probability density function f it can be calculated as follows:

$$Var[X] = \int_{\Omega} (x - E[X])^2 f(x)dx = \int_{x_-}^{x_+} (x - E[X])^2 f(x)dx. \tag{9}$$

Why do we need probability distributions instead of relying solely on sample data?

There are many reasons for using probability distributions. For coastal physical systems, the use of such functions helps us understand the fundamental physical processes that generate our data, rather than simply relying on our particular measurements, which are prone to sampling variability and irregularities ("noise"). Sample data are limited to past observations, but probability distributions extend our understanding by enabling probability estimation for unmeasured values and predictions beyond the range of our measurements. This predictive power is very important for analyzing extreme events, a topic we will address later in the course. Similarly, probability distributions help us quantify uncertainty in the statistical properties of the physical variable of interest; something that is nearly impossible using sample data alone.

The following three sections present commonly used probability distributions for modeling continuous random variables in coastal physical systems.

3.1 Normal Distribution: Sea-Surface Pressure Example

Suppose now that we have a sample containing measurements of daily average sea-surface pressure (ssP) that span one full year (2015), all recorded at the same station at Cutler Ferris Wharf, ME (Figure 1), with a barometer. The histogram of these measurements is shown in Figure 2. The distribution of our data is shown to be approximately symmetric around the mean, with roughly equal amounts of data on both sides of the mean. In other words, we can say that it resembles a symmetric "bell". A suitable probability density function (pdf) for such a sample of measurements is the so-called **normal distribution**, or else, **Gaussian distribution**. The pdf (f) for the normal distribution is given as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{10}$$

where $\mu \in R$ and $\sigma > 0$ constitute the fixed parameters of f, which are called mean and standard deviation, respectively. When our data "follow" the normal distribution, we can express it statistically by writing $X \sim N(\mu, \sigma)$ with the sample space of x being all real numbers, that is, $\Omega = \mathbb{R} = (-\infty, +\infty)$.

Let us use Wolfram Mathematica software to calculate the expectation and variance of the normal distribution by solving the definite integrals of Equations 8 and 9, respectively.

In general, our first and most important goal in statistical modeling should be to estimate the parameters of f given the data we have collected, that is,

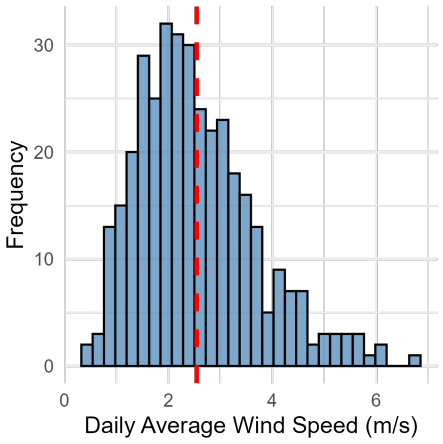

Figure 4: Histogram of daily average wind speed measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

Figure 5: Close-up of anemometer installed at the meteorological/tidal station at Cutler Ferris Wharf, ME (© NOAA Tides and Currents).

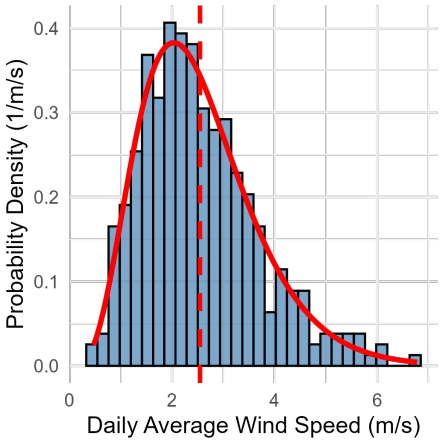

Figure 6: Probability density function of Gamma distribution (red solid line) fit to daily average wind speed measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

"fit" a distribution to our data (see **Chapter 5** later). Once we have estimated the parameters of f, we can then produce smooth density curves, such as those of Figure 3, and compute the probability that our variable falls between any two values using Equation 7.

3.2 Gamma Distribution: Wind Speed Example

We will now examine daily average wind speed measurements collected throughout 2015 from the same monitoring station at Cutler Ferris Wharf, ME, (shown in Figure 4), using an anemometer device (depicted in Figure 5). This data sample exhibits a pronounced right-skewed distribution pattern. As an approach to modeling these measurements probabilistically, we can consider using the so-called **Gamma distribution** as our probability density function (pdf). This pdf (f) can be written as:

$$f(x) = \frac{1}{\sigma^{\alpha}\Gamma(\alpha)} x^{\alpha-1} \exp\left(-x/\sigma\right), \tag{11}$$

with $\alpha > 0$ and $\sigma > 0$ being the fixed parameters of f, which are called shape and scale, respectively, while $\Gamma(\alpha)$ is defined as:

$$\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1} e^{-z} dz. \tag{12}$$

When our data "follow" the Gamma distribution, we can write it statistically as $X \sim Gamma(\alpha, \sigma)$, with the sample space of x being all positive real values, that is, $\Omega = (0, \infty)$. Therefore, the Gamma distribution is suitable for modeling continuous random variables that are skewed and take only positive values (excluding zero). Figure 6 shows a Gamma distribution fit to our daily average wind speed measurements from Figure 4.

Let us solve the respective definite integrals and calculate the expectation and variance of the Gamma distribution using Wolfram Mathematica software.

3.3 Rayleigh Distribution: Wave Height Example

An important probability density function (pdf) commonly used in coastal engineering applications is the **Rayleigh distribution**. When wave heights are measured at a specific location using a wave buoy (shown in Figure 7), these random measurements typically "follow" the Rayleigh distribution with pdf (f):

$$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{13}$$

where $\sigma > 0$ is the fixed parameter of f, known as scale. The sample space of x for the case of Rayleigh distribution is $\Omega = [0, \infty)$, that is, all positive real values (including zero). As an example, Figure 8 displays a Rayleigh pdf with a known shape parameter of $\sigma = 1.50$. This pdf was fit to wave height measurements which result to a significant wave height of $H_{1/3} \approx 3.00$ m; the histogram of these measurements is omitted from Figure 8.

Figure 7: Wave buoy at Penobscot Bay, ME, measuring incident wave heights (© NOAA National Buoy Data Center).

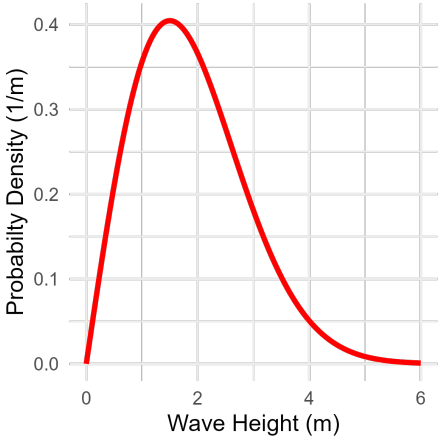

Figure 8: Probability density function of Rayleigh distribution (red solid line) with a known shape parameter of $\sigma = 1.50$.