Data Analytics for Coastal Systems

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

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

2 Random Variables and Sample Statistics

A **random variable**, usually denoted by the capital letter X, is a quantity whose outcome is uncertain. For example, the daily average temperature at a coastal location, before being measured, is a random quantity. Once measured, the average temperature on day i is now known and is denoted as x_i. A collection of measurements (x_i) is called **sample**. The set of all possible values that the sample of daily average temperature can take is called **sample space** and is denoted as Ω. In this case, $\Omega = \mathbb{R}$, since temperature can theoretically take any real value, and we say X "belongs" to $\mathbb{R}$, i.e., $X \in \mathbb{R}$.

Consider now that we have measurements of the daily average temperature for an entire year, e.g., 2020, taken from NOAA's (National Oceanic and Atmospheric Administration) meteorological station at Cutler Farris Wharf, ME (see Figure 1). A widely used form of visually summarizing such data is the **histogram** shown in Figure 2. In the histogram, the measurements of daily average temperature are divided into "bins" (categories) of equal size shown in the x-axis. While, the y-axis represents the amount of measurements (frequency) falling inside each bin. In other words, the histogram shows the **distribution** of our data.

Figure 1: Meteorological/tidal station measuring ambient temperature, among other physical variables, at Cutler Farris Wharf, ME (© NOAA Tides and Currents).

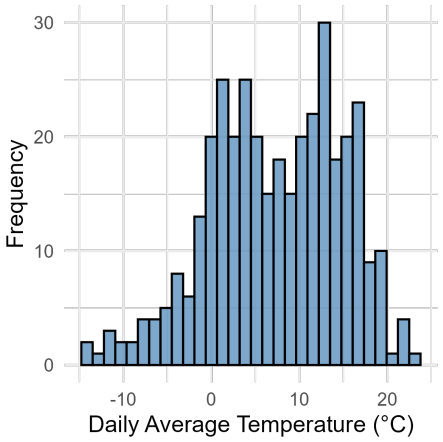

Figure 2: Histogram of daily average temperature measurements at Cutler Farris Wharf, ME, in year 2020.

Sample statistics are numerical measures derived from our sample that summarize key characteristics of the distribution of our data with a single number. The most common sample statistics are described in Sections 2.1 to 2.3 below.

2.1 Centrality

In the previous example where our random variable is the daily average temperature, in reality, the meteorological station does not collect just one measurement per day. Instead, it records temperature readings every 20 minutes and we thus derive a daily "average" by calculating the so-called arithmetic **mean ($\bar{x}$)**. This is the most common sample statistic and, arguably, the most important one. It can be computed for n measurements (x_i) as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{1}$$

While the mean is the average value of our sample, the **median ($\tilde{x}$)** is another sample statistic that represents the "middlemost" value of our data. In that sense, roughly half of the measurements in our sample are greater than or equal to this value, and roughly half are smaller than or equal to this value. Depending on whether there is an even or odd number of measurements, this sample statistic can be calculated as follows:

$$\tilde{x} = \begin{cases} x'_{(n+1)/2} & \text{for an odd sample size } n \\ (x'_{n/2} + x'_{(n+1)/2})/2 & \text{for an even sample size } n, \end{cases} \tag{2}$$

where by x'_n we denote the n^{th} value of the ordered measurements. The median is therefore the middlemost measured value (for an odd n) or the average of the two middlemost measured values (for an even n). Both the mean and the median have the same units as that of our random variable; in this case °C.

2.2 Spread

Apart from sample statistics that characterize centrality, we also need numerical measures that describe the spread of our data. The most important such sample statistic is the **standard deviation (s)**, which can be computed as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \tag{3}$$

Note that by definition, the standard deviation has the same units as that of our random variable, i.e., °C, in the case of daily average temperature.

The square of the standard deviation is called the **variance (s^2)**, which is another sample statistic for the spread of our data, and can be calculated as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \tag{4}$$

albeit it does not yield meaningful units as is the case with the standard deviation.

A ***p*-quantile** is a numerical measure that divides our sample such that approximately $100 \times p$ percent of the measurements are less than or equal to this value, and approximately $100 \times (1-p)$ percent of the measurements are greater than or equal to this value. Therefore, the median value which was introduced in Section 2.1 represents the 0.5-quantile ($q_{0.5}$). Besides the median, two other important quantiles are the 0.25- and 0.75-quantile, whose difference gives another useful sample statistic for the spread, namely, the **interquartile range (IQR)**:

$$IQR = q_{0.75} - q_{0.25}. \tag{5}$$

It is evident that the IQR has the same units as that of our random variable, i.e., °C.

2.3 Asymmetry

When the distribution of our data lacks symmetry around the mean, as is the case with the measurements of daily average temperature at Cutler Farris Wharf, ME, in year 2020 (Figure 3), then we say that the data are skewed. A sample statistic indicative of asymmetry can be calculated as follows:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}. \tag{6}$$

This value is called **skewness (g_1)** and it is unitless. When $g_1 \sim 0$, then the distribution of our data is symmetric, whereas if $g_1 > 0$ then the distribution is "right-skewed", and when $g_1 < 0$ we say it is "left-skewed".

2.4 Numerical Example

Our example comprises $n = 366$ measurements $(x_1, x_2, \ldots, x_{365}, x_{366})$ of daily average temperature at Cutler Farris Wharf, ME, in year 2020. Using Equations 1 through 6 we can compute the sample statistics as shown in Table 1.

2.5 Visualization: Boxplot

Another useful way of visually summarizing our sample, which is more descriptive than the histogram, is the so-called **boxplot**. A boxplot displays the distribution of our data using a rectangular box that shows the median (center line) and the 0.25- and 0.75-quantile (box edges). It also contains "whiskers" extending to $1.5 \times$ IQR from both $q_{0.25}$ and $q_{0.75}$. The boxplot provides a quick visual summary of a sample's center, spread, and asymmetry. The boxplot of the daily average temperature measurements at Cutler Farris Wharf, ME, in year 2020 is shown in Figure 4.

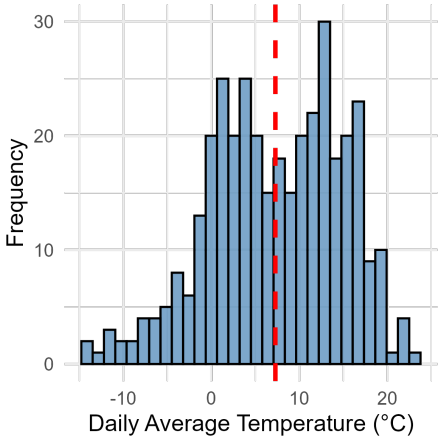

Figure 3: Histogram of daily average temperature measurements at Cutler Farris Wharf, ME, in year 2020. The mean of the data is shown with a red dashed line.

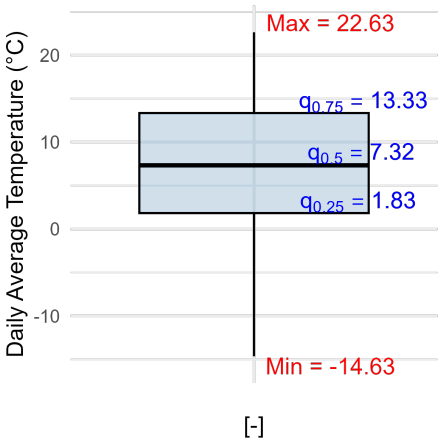

Figure 4: Boxplot of daily average temperature measurements at Cutler Farris Wharf, ME, in year 2020.

Sample Statistic	Numerical Value	R Command
Mean ($\bar{x}$)	$7.28\,(^\circ\text{C})$	**mean(x)**
Median ($\tilde{x}$)	$7.32\,(^\circ\text{C})$	**median(x)**
Standard Deviation (s)	$7.52\,(^\circ\text{C})$	**sd(x)**
Variance (s^2)	$56.55\,(^\circ\text{C})^2$	**var(x)**
Interquartile Range (IQR)	$11.50\,(^\circ\text{C})$	**IQR(x)**
Skewness (g)	-0.35 (-)	**moments::skewness(x)**

Table 1: Sample statistics for the measurements of daily average temperature at Cutler Farris Wharf, ME, in year 2020. The respective R command to compute them programmatically is also given. Note that **x** is a numeric vector in R where the measurements are stored.