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

4 Hypothesis Testing

Statistics help us to make judgments about the measurements we have collected. A powerful statistical tool that we can use for inference is the so-called "hypothesis testing" (HS). During HS, a hypothesis is proposed as an explanation for a phenomenon related to our sample of measurements. For example, such a hypothesis might be that "the daily average wind speed measurements at Cutler Farris Wharf, ME, in 2015 originate from a normal distribution". In reality, HS involves not just one but two hypotheses: the first is a **null hypothesis (H_0)**, which is the proposed explanation to be tested, and the second is an **alternative hypothesis (H_A)**, which is the statement considered as an alternative explanation to H_0. Using the same example as above, the H_A would thus be that "the daily average wind speed measurements at Cutler Farris Wharf, ME, in 2015 do not originate from a normal distribution". The idea of hypothesis testing is to check whether H_0 should be rejected in favor of H_A. The rejection criterion of H_0 involves a **test statistic**, which is a numerical quantity calculated from our sample.

When we make a decision based on HS, there is always the chance that our decision might be wrong. In fact, there is the chance that we might make what we call either a **Type I** or a **Type II** error. The table below explains these two types of errors:

	H_0 holds true	H_0 does not hold true
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Table 1: Errors associated with hypothesis testing.

Essentially, if H_0 holds true but we reject it, then this is called a Type I error, whereas if H_0 does not hold true but we do not reject it, then this is called a Type II error. The maximum acceptable probability of Type I error is usually a threshold we set as part of HS and it is called the **significance level (α)**. The goal is to compare the so-called **p-value**, which is derived using the test statistic, with the significance level α. *It is important to remember that the p-value represents the probability of obtaining the observed test statistic (or a more extreme value), given that H_0 holds true*.

In the following five steps, we summarize the general process of hypothesis testing:

1. Formulate H_0 and H_A hypotheses and set significance level (typically, $\alpha = 0.05$, or less commonly, $\alpha = 0.01$).

2. Calculate the test statistic from the sample (the test statistic depends on the hypothesis test being performed).

3. Compute the probability of obtaining the observed test statistic (or a more extreme value) given that H_0 holds true, i.e., the p-value.

4. Compare p-value ("one-sided") or "$2 \times$ p-value" ("two-sided") with α.

5. Do not reject H_0 if $(2\times)$p-value $> \alpha$, otherwise reject.

The next three sections cover useful hypothesis tests that allow us to: 1) evaluate differences between the distributions of independent samples (e.g., seasonal measurements), 2) assess whether a monotonic trend exists in our measurements over time, and 3) check whether our data follow a particular distribution (namely, the normal distribution),

4.1 Mann-Whitney U Test

The use of this test is for detecting differences in the distributions of two independent samples. The null hypothesis here is that there are no differences between the distributions of the two samples, while the alternative hypothesis is that there are indeed differences. Another interpretation of the Mann-Whitney U test is that it checks for differences in the medians of the two distributions, with the null hypothesis being that the two medians are equal and the alternative hypothesis that the median of the first sample is either lower or higher than that of the second sample. Let us assume that $x_1, x_2, \ldots, x_{n_1}$ are independent and identically distributed (*i.i.d.*) measurements of the first sample and $y_1, y_2, \ldots, y_{n_2}$ are *i.i.d.* measurements of the second sample with sizes n_1 and n_2, respectively. Then, the Mann-Whitney U test statistic (U) can be computed as follows:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \quad (1)$$

where R_1 is computed as the sum of the ranks of the measurements in the first sample, after we pool the measurements from both samples and rank them from smallest to largest. The test statistic is approximately normally distributed for relatively large sample sizes ($n_1 \geq 20, n_2 \geq 20$) with mean and variance:

$$\mu = \frac{n_1 n_2}{2}, \quad \sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad \text{(assuming no data ties).} \quad (2)$$

Hence, we can obtain the p-value of U using the normal distribution. The null hypothesis of no differences in the distributions of the two samples is rejected for small (extreme) values of U, i.e., for a very small p-value. We summarize the Mann-Whitney U test below:

1. H_0 : no differences exists VS. H_A : differences exists.
2. Calculate U based on Equation 1.
3. Compute the p-value from the respective normal distribution (Equation 2).
4. Compare "$2 \times$ p-value" (account for either lower or higher median, i.e., "two-sided") with 0.05.
5. There are no differences in the distributions of the two samples if $2 \times$ p-value > 0.05.

Let us use R and the **wilcox.test()** function to perform a Mann-Whitney U test and check for differences in the distributions of the seasonal daily average wind speed measurements (Figure 1).

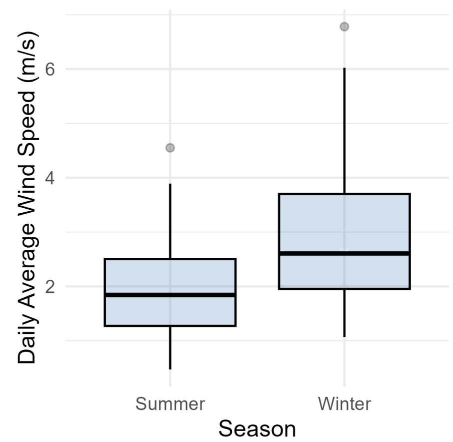

Figure 1: Boxplots of seasonal daily average wind speed measurements at Cutler Farris Wharf, ME, in year 2015.

4.2 Mann-Kendall Test

This test is useful for detecting monotonic trends (increasing/decreasing) in our data over time and works well for $n \geq 8$, where n is the sample size. It tests the null hypothesis that no such trend exists, against the alternative hypothesis that either a negative or a positive trend is present in our measurements over time. Assuming *i.i.d.* measurements $(x_1, x_2, \ldots, x_n)$, the Mann-Kendall test statistic (S) is expressed as:

$$S = \Sigma_{i=1}^{n-1} \Sigma_{j=i+1}^{n} \text{sgn}(x_j - x_i), \quad (3)$$

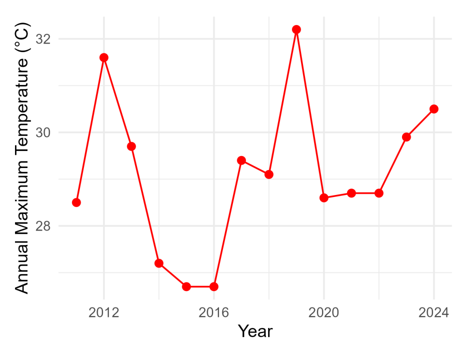

Figure 2: Time series of annual maximum temperature measurements at Cutler Farris Wharf, ME.

where x_j and x_i are latter and earlier (in time) measurements of our sample, respectively, while sgn is a function given by:

$$\text{sgn}(\zeta) = \begin{cases} -1, & \zeta < 0 \\ 0, & \zeta = 0 \\ 1, & \zeta > 0. \end{cases}$$

The test statistic is approximately normally distributed with the following mean and variance:

$$\mu = 0, \quad \sigma^2 = \frac{n(n-1)(2n+5)}{18} \quad \text{(assuming no data ties).} \tag{4}$$

Therefore, we can compute the p-value of S using the normal distribution. The null hypothesis of no monotonic trend is rejected for large (extreme) values of $|S|$, i.e., for a very small p-value. The Mann-Kendall test can be summarized as follows:

1. H_0 : no monotonic trend exists VS. H_A : monotonic trend exists.
2. Calculate S based on Equation 3.
3. Compute the p-value from the respective normal distribution (Equation 4).
4. Compare "$2 \times$ p-value" (account for either positive or negative trend, i.e., "two-sided") with 0.05.
5. There is a monotonic trend in our sample measurements if $2 \times$ p-value ≤ 0.05. Note: the trend is positive if $S > 0$, or else is negative for $S < 0$.

We can now use R and the **Kendall::MannKendall()** function to perform a Mann-Kendall test and check for a monotonic trend in annual maximum temperature measurements (Figure 2).

4.3 Shapiro-Wilk Test

This test can be used to assess the normality of a sample's distribution and is better suited for $20 \leq n \leq 2000$, where n is the sample size. In other words, it tests the null hypothesis that our measurements come from a normal distribution, against the alternative hypothesis that they do not. If our sample is composed of n measurements which are *i.i.d.* $(x_1, x_2, \ldots, x_n)$, and if $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ represent the measurements in ascending order, then the Shapiro-Wilk test statistic (W) is given by:

$$W = \frac{(\Sigma_{i=1}^{n} \alpha_i x_{[i]})^2}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}, \tag{5}$$

where $\bar{x}$ is the sample mean and α_i's are constants derived as functions of n. The p-value associated with W can be calculated by first transforming W so that it is normally distributed under the null hypothesis. The null hypothesis of normality of our data is rejected for small (extreme) values of W, i.e., for a small p-value (≤ 0.05). The Shapiro-Wilk test can be summarized as follows:

1. $H_0 : X \sim N(\mu, \sigma)$ VS. $H_A : X \nsim N(\mu, \sigma)$.
2. Calculate W based on Equation 1.
3. Transform W to make it normally distributed and compute the respective p-value.

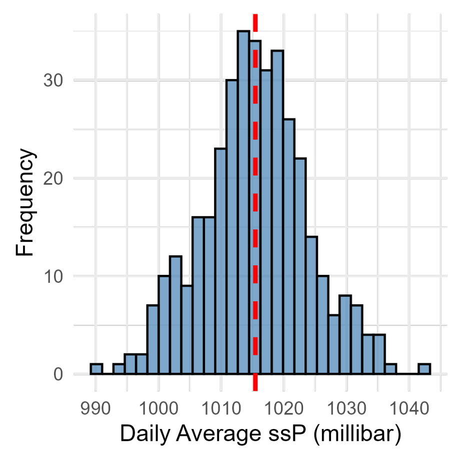

Figure 3: Histogram of daily average sea-surface pressure (ssP) measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

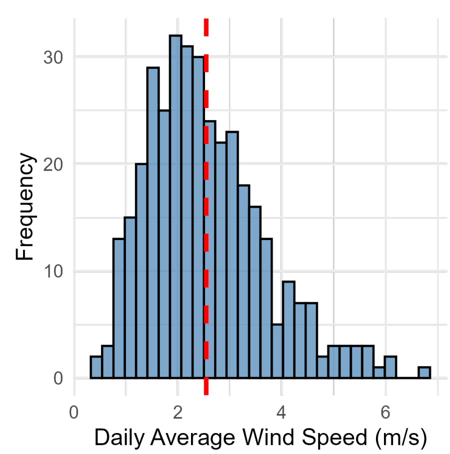

Figure 4: Histogram of daily average wind speed measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line.

4. Compare p-value with 0.05 (i.e., "one-sided").

5. Our sample measurements originate from a normal distribution if p-value $>$ 0.05.

With the use of R and the **shapiro.test()** function let's perform a Shapiro-Wilk test and check for normality of the daily average wind speed and sea-surface pressure measurements (Figures 3 and 4).

For more details about the Shapiro-Wilk test (e.g., on how to obtain α_i's), refer to Royston (1982) and Royston (1995).

References

1) Royston, J. Patrick. "An extension of Shapiro and Wilk's W test for normality to large samples." Journal of the Royal Statistical Society: Series C (Applied Statistics) 31, no. 2 (1982): 115-124.

2) Royston, Patrick. "Remark AS R94: A remark on algorithm AS 181: The W-test for normality." Journal of the Royal Statistical Society. Series C (Applied Statistics) 44, no. 4 (1995): 547-551.