The Time Domain

The Sample Interval

We have a sequence of N observations

$$z_n, \qquad n = 0, 1, 2, \ldots N - 1$$

which coincide with times

$$t_n, \qquad n = 0, 1, 2, \ldots N - 1.$$

The sequence z_n is called a *discrete time series*.

It is assumed that the *sample interval*, Δ, is constant

$$t_n = n\Delta$$

with the time at $n = 0$ defined to be 0. The *duration* is $T = N\Delta$.

If the sample interval in your data is not uniform, the first processing step is to interpolate it be so.

The Underlying Process

A critical assumption is that there exists some “process” $z(t)$ that our data sequence z_n is a *sample of*:

$$z_n = z(n\Delta), \qquad n = 0, 1, 2, \ldots N-1.$$

Unlike z_n, $z(t)$ is believed to exist for *all times*.

(i) The process $z(t)$ exists in *continuous time*, while z_n only exists at *discrete times*.

(ii) The process $z(t)$ exists for *all past and future* times, while z_n is only available over a certain time interval.

It is the properties of $z(t)$ that we are trying to estimate, *based on* the available sample z_n.

Measurement Noise

In reality, the measurement device and/or data processing probably introduces some artifical variability, termed *noise*.

It is more realistic to consider that the observations contain samples of the process of interest, $z(t)$, *plus* some noise ϵ_n:

$$z_n = z(n\Delta) + \epsilon_n, \qquad n = 0, 1, 2, \ldots N - 1.$$

This is an example of the *unobserved components model*. This means that we *believe* that the data is composed of *different components*, but we cannot observe these components individually.

The process $z(t)$ is potentially obscured or degraded by the limitations of data collection in three ways: (i) finite sample interval, (ii) finite duration, (iii) noise.

Because of this, the time series is an *imperfect* representation of the real-world processes we are trying to study.

A Pair of Time Series

In oceanography we often have a *pair* of time series x_n and y_n. Such data is called *bivariate*, meaning that it is consists of two variables.

These may represent horizontal velocity (as in current meters) or displacement (floats or drifters).

Bivariate data can be thought of as a vector having two elements:

$$\mathbf{z}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}.$$

The subscript n here refers to n different copies of the vector, *not* to the elements of that vector!

Alternatively, we can also think of this data consisting of a single *complex-valued* time series $z_n \equiv x_n + iy_n$, where $i \equiv \sqrt{-1}$.

Vector and complex notations will both be discussed in detail later.

Time versus Frequency

There are two opposite points of view regarding the time series z_n.

The first regards z_n as being built up as a sequence of discrete values $z_0, z_2, \ldots z_{N-1}$.

This is the domain of *statistics*: the mean, variance, histogram, etc.

When we look at data statistics, generally, the order in which the values are observed *doesn't matter*.

The second point of view regards z_n as being built up of sinusoids: purely periodic functions spanning the whole duration of the data.

This is the domain of *Fourier spectral analysis*.

In between these two extremes is wavelet analysis.

This lecture will focus on what can be done in the time domain.

Time-Domain Statistics

A good place to start is with the very simplest tools.

The *sample mean* describes a “typical” value:

$$\overline{z} \equiv \frac{1}{N} \sum_{n=0}^{N-1} z_n$$

The *sample variance* gives the spread about the mean:

$$\sigma_z^2 \equiv \frac{1}{N} \sum_{n=0}^{N-1} \left(z_n - \overline{z}\right)^2$$

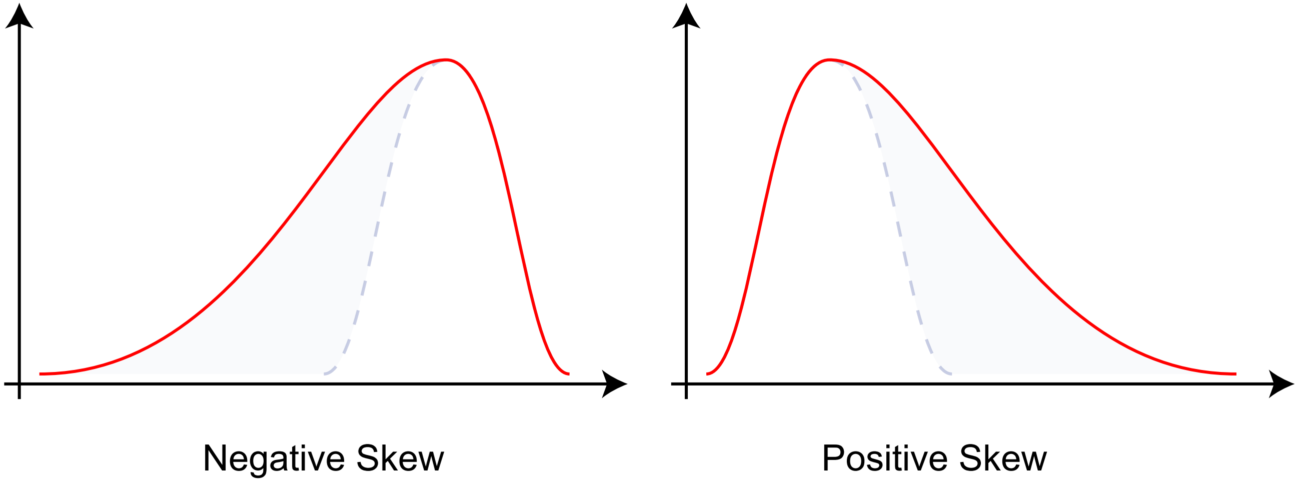
“Sample” here means that it is computed from the observed data, as opposed to being a property of the assumed underlying process $z(t)$.

The mean and variance are called the first two *moments* of the distribution of values associated with the process.

Skewness

The *skewness* describes the tendency for an *asymmetry* between positive excursions and negative excursions:

$$\gamma_z \equiv \frac{1}{\sigma_z^3} \frac{1}{N} \sum_{n=0}^{N-1} \left(z_n - \bar{z}\right)^3$$



Kurtosis

The *kurtosis* is said to either measure *peakedness* (concentration near $\overline{z}$), or a tendency for *long tails* (concentration far from $\overline{z}$):

$$\kappa_z \equiv \frac{1}{\sigma_z^4}\frac{1}{N}\sum_{n=0}^{N-1}\left(z_n - \overline{z}\right)^4$$

Actually, it measures both. Kurtosis is a measure of the spread of z_n about the *two points* $\overline{z} \pm \sigma_z$. This can happen *either* for peakness *or* for long tails! *See Moors (1986), "The Meaning of Kurtosis".*

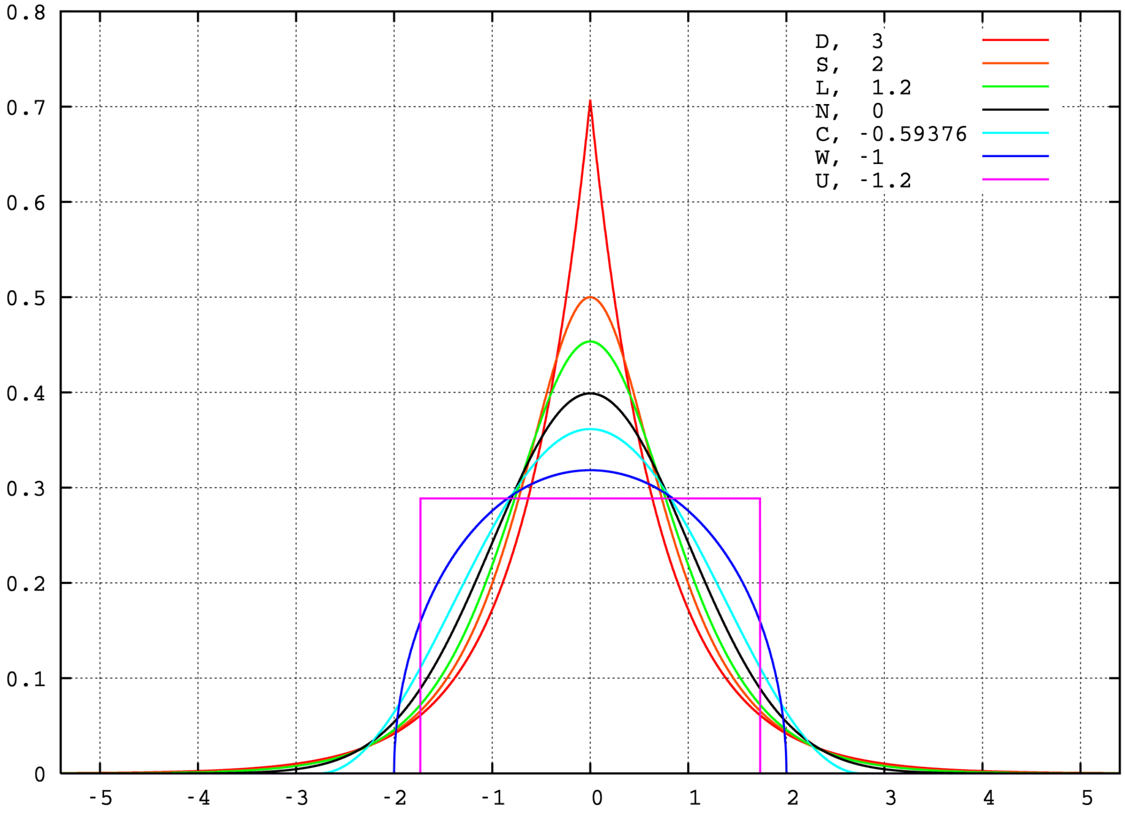
Because the value of kurtosis for a Gaussian process can be shown to be equal to 3, one sometimes encounters the *excess kurtosis*

$$\tilde{\kappa}_z \equiv \kappa_z - 3.$$

Values of $\tilde{\kappa}_z > 0$ mean the data is *more kurtotic*—peaked or long-tailed—than a Gaussian, while $\tilde{\kappa}_z < 0$ means it is less so.

Illustration of Kurtosis

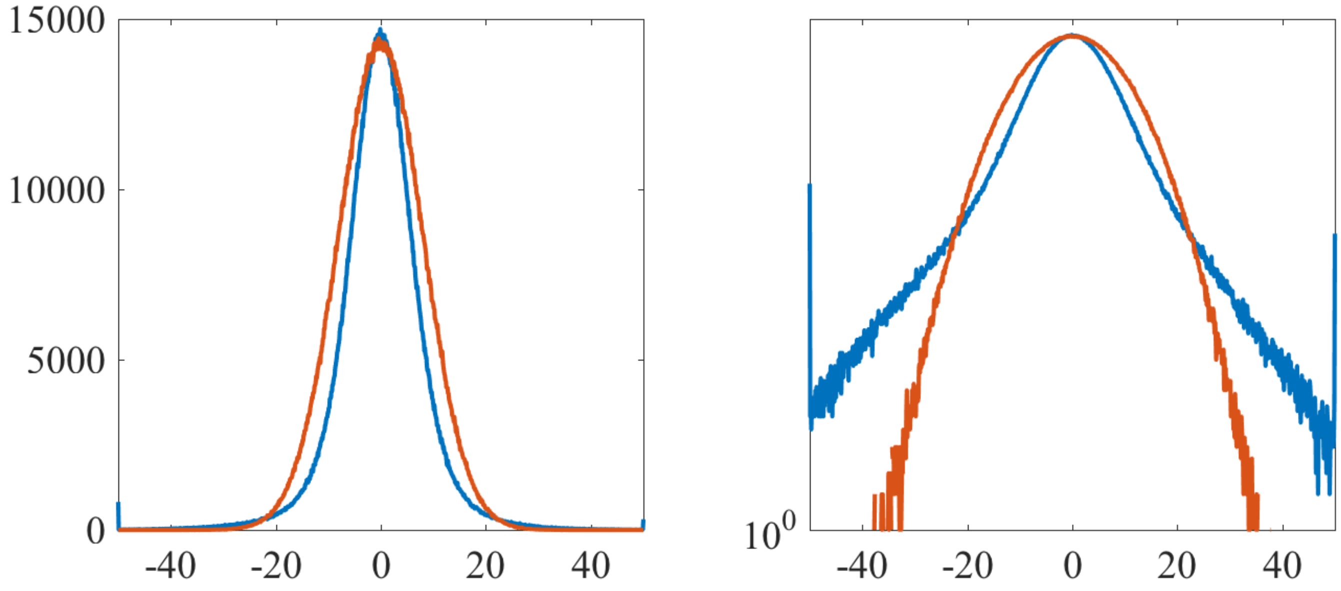
Distributions corresponding to different values of excess kurtosis.



Positive excess kurtosis corresponds to long tails and peakedness.

Histogram

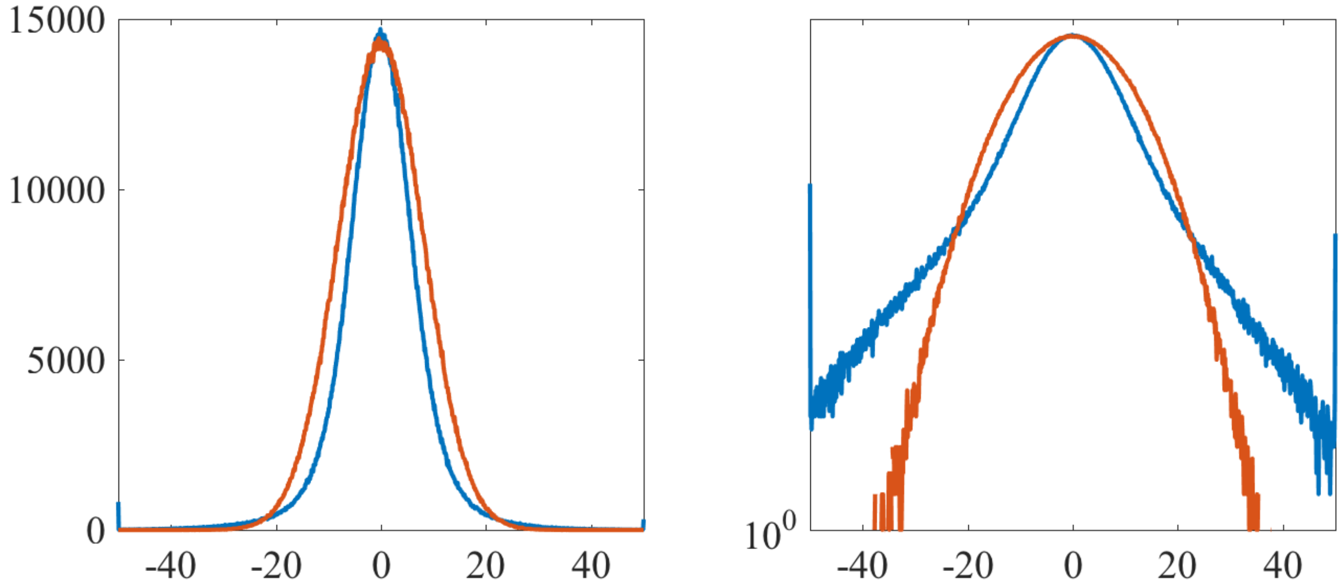
The mean, variance, skewness, and kurtosis describe aspects of the *histogram*: the observed distribution of data values.



Here is the histogram of *all* SSH values from long altimeter track (blue), versus Gaussian noise having the same variance (orange).

Histogram

The mean, variance, skewness, and kurtosis describe aspects of the *histogram*: the observed distribution of data values.



Here is the histogram of *all* SSH values from long altimeter track (blue), versus Gaussian noise having the same variance (orange).

Q: Is it skewed? Is it kurtotic?

Simple Smoothing

One of the most effective ways to process a time series is with a simple smoothing.

Let g_m be a length M sequence, where M is *odd*, defined for

$$-(M-1)/2, \ldots, -2, -1, 0, 1, 2, \ldots, (M-1)/2.$$

Note that we define g_m to be centered on $m = 0$, instead of running between 0 and $M - 1$.

A *smoothed* version of the discrete time series z_n is defined as

$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m$$

where g_m is called the *filter* or the *smoothing window*. It is also useful to examine the *residuals* from the original, $\breve{z}_n \equiv z_n - \tilde{z}_n$.

Simple Smoothing Example

An example of simple smoothing is a *running mean*. A five-point running mean is given by:

$$\tilde{z}_n = \frac{1}{5}\left[z_{n-2} + z_{n-1} + z_n + z_{n+1} + z_{n+2}\right] .$$

This is expressed by the filtration equation

$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m$$

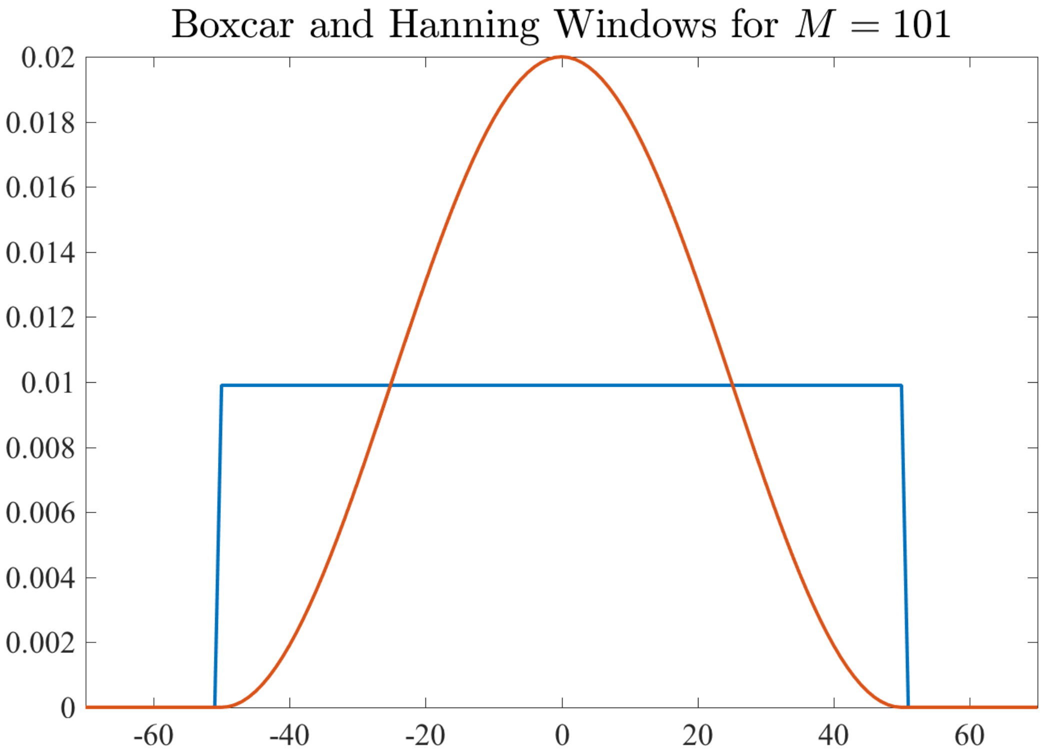
with the choice

$$g_m = 1/5, \qquad m = -2, -1, 0, 1, 2.$$

The simplest choice of filter is $g_m = 1/M$, a constant over the M points. Then the filtration defines an M-point running mean.

Choice of Filter

The running mean filter $g_m = 1/M$ is called the *boxcar* or *rectangle function*. Another popular choice is the *Hanning window*.



The Hanning window is just a half-period of a cosine, offset.

How to Choose a Filter

The goal of simple smoothing is to separate relatively “fast” from relatively “slow” variability.

Many functions can be used as smoothing filters. However, for a first look at the data, the details of the filter are not so important.

The important thing is to define a sensible *weighted average*.

The boxcar filter has sharp “edges” that can lead to artifacts, as we will see later. Also, the boxcar is highly distributed, and doesn't place emphasis on the “present time” compared to nearby times.

For these reasons, the Hanning window is sometimes more appropriate for simple smoothing.

In `jLab` simple smoothing is carried out with `vfilt`.

How to Choose a Filter

The goal of simple smoothing is to separate relatively “fast” from relatively “slow” variability.

Many functions can be used as smoothing filters. However, for a first look at the data, the details of the filter are not so important.

The important thing is to define a sensible *weighted average*.

The boxcar filter has sharp “edges” that can lead to artifacts, as we will see later. Also, the boxcar is highly distributed, and doesn't place emphasis on the “present time” compared to nearby times.

For these reasons, the Hanning window is sometimes more appropriate for simple smoothing.

In `jLab` simple smoothing is carried out with `vfilt`.

Q: Why don't we use a Gaussian?

What to do at Endpoints?

Smoothing runs into a difficulty near the endpoints of z_n:

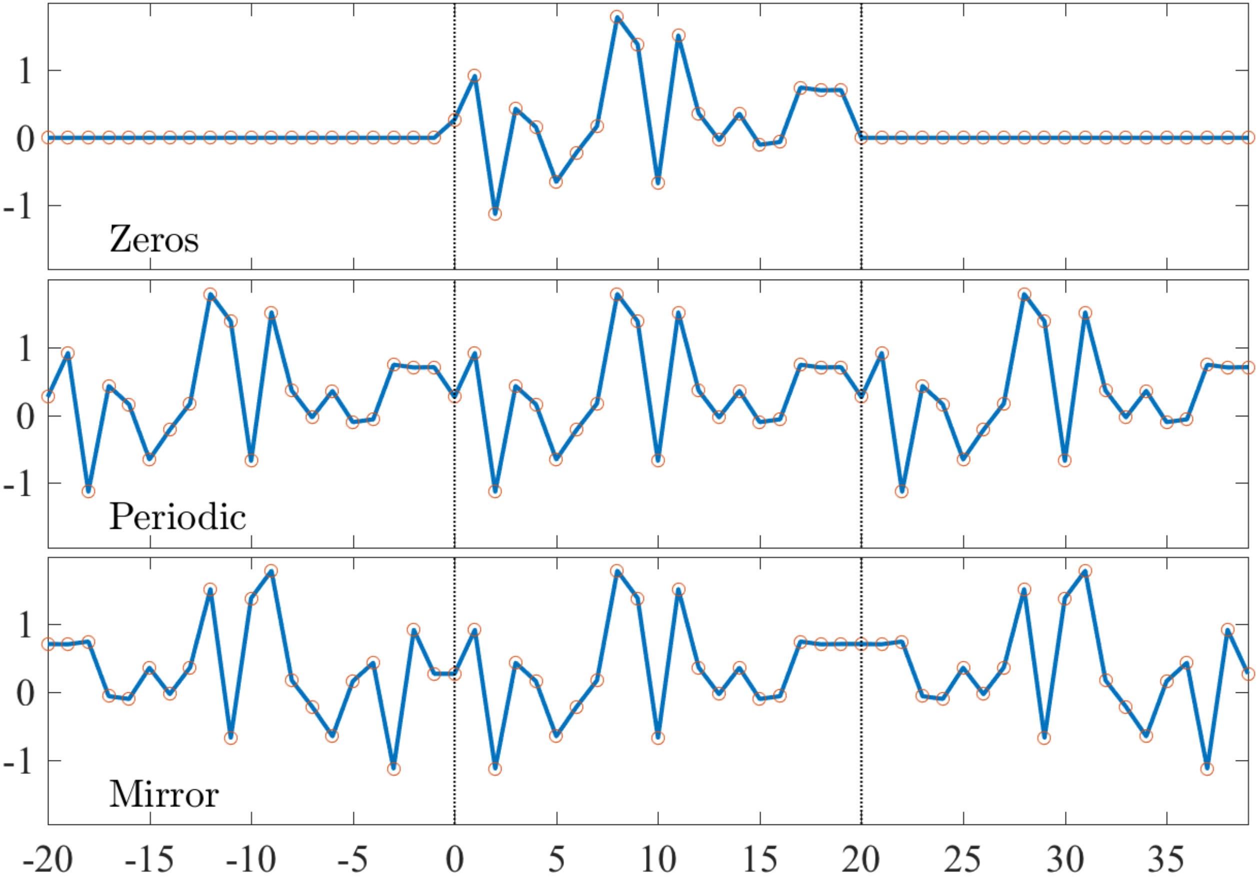
$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m .$$

When we are within a filter half-width $(M-1)/2$ of the beginning or end of z_n, the filter "falls off" the end of the data.

Some choice must be made in order to have the smoothed version $\tilde{z}_n$ of the data be well defined. There are five common choices.

1. **Truncate**: Omit affected points, such that the length of $\tilde{z}_n$ will be about M points *less than* the length of z_n.
2. **NaNs**: Replace these with NaNs or *indeterminate* values.
3. **Zeros**: Set z_n equal to zero for $n \leq 0$ or $n \geq N-1$.
4. **Periodic**: Make z_n periodic by wrapping around the ends.
5. **Mirror**: Reflect z_n about its beginning and also about its end.

Endpoint Illustration



The *mirror* condition generally leads to the fewest “edge effects”, especially when the data is nonstationary or has a linear trend.

Summary

This lecture has focused on

Summary

This lecture has focused on

- Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

Summary

This lecture has focused on

- Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

- Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

Summary

This lecture has focused on

- Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

- Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

- Discussing *simple smoothing* and details of its implementation.

Summary

This lecture has focused on

- Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

- Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

- Discussing *simple smoothing* and details of its implementation.

My experience is that *looking at data* together with *statistics* and *simple smoothing* is maybe 50% of analyzing time series!

There are more sophisticated tools that can often, but not always, be very useful in unlocking the potential of the data.

However, learning how to make use of these takes a lot more work!

To be continued...

In-Class Assignments

1. Compute and plot the histogram of your data. You can do this using Matlab's `hist` or `histogram` functions. (For bivariate data, do this and the next step for the both components.)
2. Compute the sample mean, variance, skewness, and kurtosis using Matlab's `mean` and `std` functions. You can also compute these quantities from the histrogram your made earlier using jLab's `pdfprops`. Do these match exactly? If not, why not?
3. Repeat 1&2 for a realization of Gaussian white noise using `randn` that is set to have the same variance as your data. In what ways, if any, is your data non-Gaussian?
4. Experiment with filtering your data and `vfilt`. Plot the data, the filtered version, and the residual (original minus filtered) for a few choices of filter length. What choice seems most suitable for your data and why? Note if your data doesn't have noise or multiple scales of varability, try working with {this one}.
5. Re-do the steps 1&2 involving the time-domain statistics, but using firstly the smoothed, and secondly the residual, versions of your data. How do the statistics change dependent upon the lowpass or highpass filtering? How do you interpret this?