Lecture 2: Probing the power of probability density functions

Reading: Bendat and Piersol, Ch. 3.1-3.3

Announcement: Field trip on Friday, October 16, 9:30 am. We'll meet at the base of the pier.

Recap

Lecture 1 adddressed some basic statistical measures. We looked at examples of time series, and we considered some basic terminology. If we have a collection of data (*random variables*), we can compute their *mean*, *variance*, *standard deviation*, *median*, and we can examine the *probability density function* of the data. We also made a distinction between the *expectation value* (the value we'd expect if we had an infinite number of perfectly sampled observations) and the observed mean. Similarly, we can distinguish between an empirical probability density function (what we actually observe) and the idea probability density function that we observe.

An example

Melissa Carter, the Shore Stations Program manager, is in charge of the pier data. When we visit the pier on October 16 (for those who are available), she'll tell us in detail about the data collection system. The system includes automated sensors that suffer from biofouling and manual time series that are under sampled. Melissa asks, "What new information is gained with the continuous 4min time series, and is there a need to continue to collect the manual once per day measurements?"

We could start to answer these questions by using the tools we reviewed in the first lecture. Are the means the same? Are the variances the same? But that will give us an incomplete picture for several reasons:

1. As we noted last time, data sets can perversely have the same mean and standard deviation, but have pdfs that look nothing alike.

2. When we deal with real data, nothing is ever identical, so we'll need to know how big a difference is acceptable.

The pdf is going to help us work through these issues. What can we do with a pdf? Let's cover three topics:

1. How do we define a pdf?

2. How do we use the pdf to think about confidence limits? Are two estimates different?

3. How can we tell if two pdfs are different?

The formal definition of a probability density function is based on the first derivative of the probability:

$$p(x) = \lim_{\Delta x \to 0} \left[\frac{\text{Prob}[x < x(k) \leq x + \Delta x]}{\Delta x} \right] \tag{1}$$

where $x(k)$ is a random variable. This definition means that the integral of the probability density function gives us the probability, as we noted last time:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{2}$$

One of the clever aspects of the pdf is that we can use it to determine an expected value:

$$E(x(k)) = \int_{-\infty}^{\infty} xp(x)\,dx = \mu_x. \tag{3}$$

Why does this work? In essence, I reorder all the values in my data set and ask what's the probability of finding x in bin 1, what's the probability of finding x in bin 2, etc? Or in other words, what fraction of my total record is in bin 1, what fraction is in bin 2, etc? And summing this way, I'll find the mean.

We can also use this for x^2 or for $(x - \mu_x)^2$.

$$E((x(k) - \mu_k)^2) = \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x)\,dx = \sigma_x^2. \tag{4}$$

Let's start by making an empirical pdf. Last time we talked about temperature on the pier, so now, let's take a look at pressure. We could plot a histogram of the data using the hist function, but that wouldn't give us a pdf. For the pdf we need to be properly normalized. We can still do this using hist:

```
% read the data
time=...
  ncread('http://sccoos.org/thredds/dodsC/autoss/scripps_pier-2019.nc',...
  'time');
pressure=...
  ncread('http://sccoos.org/thredds/dodsC/autoss/scripps_pier-2019.nc',...
  'pressure');

%time=ncread('scripps_pier-2019.nc','time');
%pressure=ncread('scripps_pier-2019.nc','pressure');
%
% compute the histogram
dx1=.1;
[a,b]=hist(pressure,2:dx1:5);
plot(b,a/sum(a)/dx1,'LineWidth',3)
%
% or using a different dx
dx2=0.01
hold on
[c,d]=hist(pressure,2:dx2:5);
plot(d,c/sum(c)/dx2,'r','LineWidth',2)
xlabel('pressure (dbar)','FontSize',14)
ylabel('probability density','FontSize',14)
set(gca,'FontSize',14)
```

Rob Pinkel, who taught this class before I took over, always told students that they couldn't use the "hist" function for this and should do a loop. How do we do that?

```
% or using a loop
dx=0.1; clear n_bin;
bins=2:dx:5;
```

```
for i=1:length(bins)
 n_bin(i)=length(find(pressure>bins(i)-dx/2 & pressure<=bins(i)+dx/2));
end
plot(bins,n_bin/sum(n_bin)/dx,'g')
```

We like to think that geophysical variables are normally distributed, meaning that the distribution is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5}$$

where μ is the mean and σ is the standard deviation. So we can add a Gaussian to our plot:

```
sigma=std(pressure); mu=mean(pressure)
plot(bins,1/sigma/sqrt(2*pi)*exp(-(bins-mu).^2 /(2*sigma^2)),...
   'k','LineWidth',2)
```

See Figure 1 for the results.

We like the Gaussian, because it's easy to calculate, and it has well defined properties. We know that 68% of measurements will be within $\pm\sigma$ of the mean, and 95% of measurements will be within $\pm 2\sigma$ of the mean.

We can turn this around to decide whether a measurement is an outlier. If we expect to see a lot of values near the mean, and we find that we have a measurement that deviates from the mean by 5σ, then it's not terribly statistically likely. (For a Gaussian, 99.99994% of observations should be within $\pm 5\sigma$ of the mean.) Thus we might decide to throw out all outliers that differ from the mean by more than 3 or 4 or 5σ.

We can also use this framework to think about uncertainty. If we measure one realization of an estimate of the mean, that will become our best estimate of the mean. If our formal estimate of our a priori uncertainty is correct (and we might also call this σ, but let's use δ for now), then we expect that 68% of the time, our single observation should be within $\pm\delta$ of the true value, and 95% of the time, our single observation should be within $\pm 2\delta$ of the true value.

Later in this class, we'll come to the concept of convolution (which is effectively a filter). One reason that we really like the Gaussian is because the convolution of a Gaussian with another Gaussian is still a Gaussian, so we can manipulate the statistics easily. But are data necessarily normally distributed?

So this might lead you to think that all data are fairly Gaussian.

Example pdfs of real data? Non-Gaussian cases.

Now what if we plot chlorophyll?

```
hold off
chl=...
  ncread('http://sccoos.org/thredds/dodsC/autoss/scripps_pier-2019.nc',...
  'chlorophyll');
flag=...
  ncread('http://sccoos.org/thredds/dodsC/autoss/scripps_pier-2019.nc',...
  'chlorophyll_flagPrimary');
%chl=ncread('scripps_pier-2019.nc','chlorophyll');
%flag=ncread('scripps_pier-2019.nc','chlorophyll_flagPrimary');
xx=find(flag==1);
```

```
[a,b]=hist(chl(xx),.5:49.5);
plot(b,a/sum(a),'LineWidth',2)
hold on
ylabel('probability density','FontSize',14)
xlabel('chlorophyll (\mu g/L)','FontSize',14)
%
mu=mean(chl(xx));
sigma=std(chl(xx));
bins=.5:49.5;
plot(bins,1/sigma/sqrt(2*pi)*exp(-(bins-mu).^2 /(2*sigma^2)),...
    'k','LineWidth',2)
set(gca,'FontSize',14)
```

As illustrated in Figure 2, chlorophyll concentrations are decidedly non-Gaussian. (We usually refer to chlorophyll as being log-normally distributed, meaning that the log of the values might be Gaussian.)

Ocean velocity data often have a double-exponential distribution, as do wind velocity data:

$$p(x) = \frac{1}{\sigma\sqrt{2}} \exp\left[-\frac{|x|\sqrt{2}}{\sigma}\right]. \tag{6}$$

Sometimes we only measure wind speed, and that's necessarily positive. The Rayleigh distribution is sometimes a good representation of wind speed: it is defined from the square root sum of two independent Gaussian components squared, $y = \sqrt{x_1^2 + x_2^2}$.

$$p(y) = \frac{y}{\sigma^2} \exp\left[-\frac{y^2}{2\sigma^2}\right]. \tag{7}$$

Summing variables, error propagation, and the central limit theorem

Given that so many pdfs can be non-Gaussian, why do we spend so much time talking about Gaussians? There are two important reasons.

1. As noted above, the Gaussian is mathematically tractable.

2. Even though individual pdfs are non-Gaussian, if we sum enough variables, everything is Gaussian. (This is the central limit theorem, which we'll get to next time.)

Often the quantities we study represent a summation of multiple random variables. For example, we're not interested in the instantaneous temperature but the average over an hour or a day. Thus we consider

$$x(k) = \sum_{i=1}^{N} a_i x_i(k), \tag{8}$$

following the terminology of Bendat and Piersol, where a_i is a coefficient. The mean of x is

$$\mu_x = E(x(k)) = E\left[\sum_{i=1}^{N} a_i x_i(k)\right] = \left[\sum_{i=1}^{N} a_i E(x_i(k))\right] = \sum_{i=1}^{N} a_i \mu_i. \tag{9}$$

and

$$\sigma_x^2 = E\left[(x(k) - \mu_x)^2\right] = E\left[\sum_{i=1}^{N} a_i (x_i(k) - \mu_i)\right]^2 = \sum_{i=1}^{N} a_i^2 \sigma_i^2. \tag{10}$$

In doing this, we've carried out a little sleight of hand, by assuming that for a large ensemble (as the number of elements used to define our expectation value E approaches ∞) the correlation between x_i and x_j is zero so that the expectation value $E[(x_i(k) - \mu_i)(x_j(k) - \mu_j)] = 0$ for $i \neq j$.

This gives us some simple rules of thumb:

Standard error of the mean. Suppose that a_i is an averaging operator and is equal to $1/N$, and σ_i is the same for all i. Then

$$\sigma_x^2 = \sum_{i=1}^{N} a_i^2 \sigma_i^2 = \frac{N\sigma_i^2}{N^2} = \frac{\sigma_i^2}{N}. \tag{11}$$

This means that the standard deviation of the mean, *the standard error of the mean*, is $\sigma/\sqrt{N}$.

As a footnote to this, the *standard error of the variance* is $\sigma^2\sqrt{2/(N-1)}$.

Error Propagation Our consideration of the summed variables gives us a rule for estimating uncertainties of computed quantities. If we sum a variety of measures together, then the overall uncertainty will be determined by the square root of the sum of the squares:

$$\delta_y = \sqrt{\sum_{i=1}^{N} a_i^2 \delta_i^2}, \tag{12}$$

where here we're using δ_i to represent the a priori uncertainties.

What if we have to multiply quantities together? Then we simply linearize about the value of interest. We'll do this properly next time. So if $y = x^2$, and we have an estimate of the uncertainty in x, δ_x, then we know that locally, near x_o, we can expand in a Taylor series:

$$y(x_o + \Delta x) = y(x_o) + \frac{dy}{dx}\Delta x. \tag{13}$$

This means that I can use my rules for addition to estimate the uncertainty in y:

$$\delta_y(x_o) = \left|\frac{dy(x_o)}{dx}\right| \delta_x = 2x_o\delta_x \tag{14}$$

and you can extend from here. If $y = a_1x + a_2x^2 + a_3x^3$, what is δ_y? When will this estiamte of uncertainty break down?

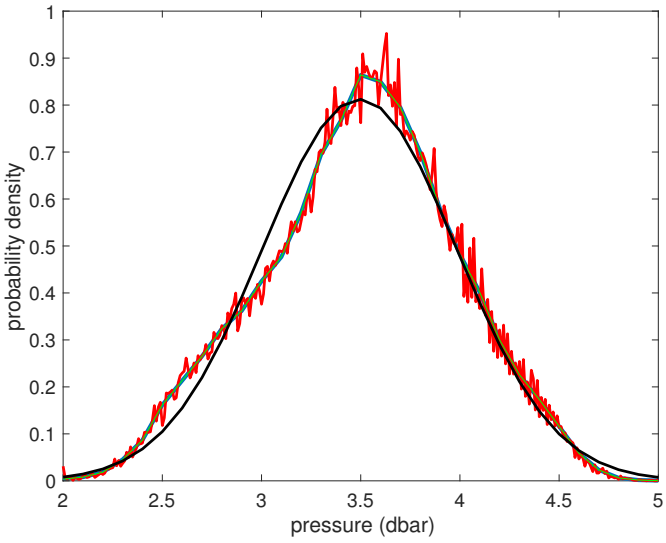

Figure 1: Probability density function for 2019 pressures measured from the shore station at the Scripps pier. Here the blue line indicates the pdf computed using bins with a width of 0.1 dbars, and the red line indicates the pdf for bins with a width of 0.01 dbars. The black line is a Gaussian defined by the mean and standard deviation of the measurements.

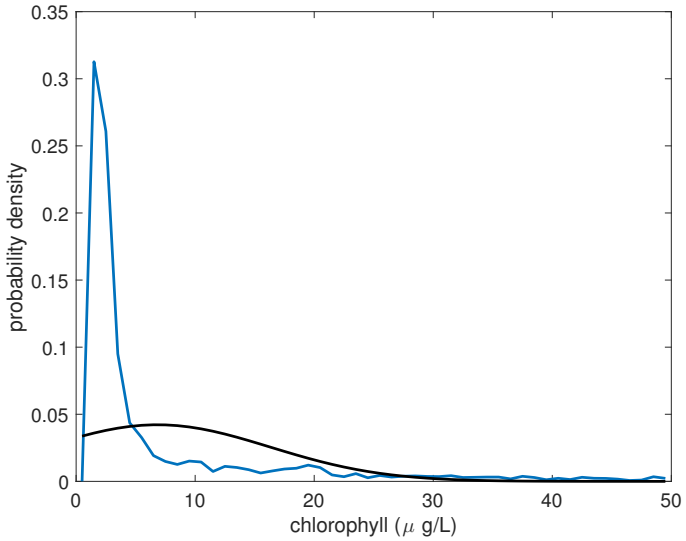

Figure 2: Probability density function for 2019 chlorophyll measured from the shore station at the Scripps pier. Here the blue line indicates the empirical pdf, and the black line is a Gaussian defined by the mean and standard deviation of the measurements.