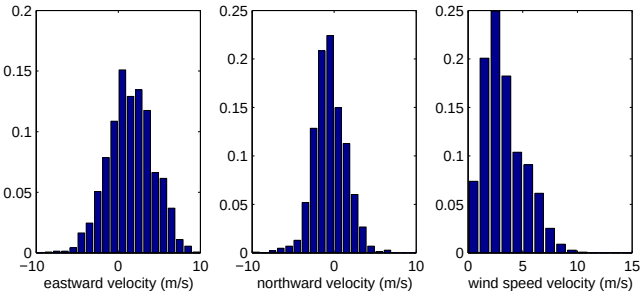

Figure 1: Empirical probability density functions for (left) eastward wind velocity, (center) northward wind velocity, (right) wind speed from the National Centers for Environmental Prediction reanalysis for the year 2000 for a grid point located approximately at San Diego.

Probability Density Functions

Histograms are easy to plot, but they aren't universal in character, so if we want to take a more general view of our data, we need to plot the probability density function or PDF.

Formal Definitions

PDFs tells us the probability of observing a value within a specific range. If P is the PDF, $P(x)dx$ is the probability of observing a value between x and $x + dx$. This notation can be a little confusing, but it has several important features. The pdf has no dependence on bin width or total sample size. It lets us determine the probability of observing a value in any arbitrary range:

$$\text{Prob}[x_1 < x < x_2] = \int_{x_1}^{x_2} P(x)dx. \tag{1}$$

Accordingly, the probability of observing a value x between $-\infty$ and $+\infty$ is clearly 100% or 1. Thus,

$$\text{Prob}[-\infty < x < \infty] = \int_{-\infty}^{\infty} P(x)dx = 1. \tag{2}$$

The *cumulative distribution function* $C(x)$ is the probability of observing a value less than x. It can be computed by integrating the pdf.

$$C(x) = \int_{-\infty}^{x} P(x')dx'. \tag{3}$$

$C(x)$ is 0 when x approaches minus infinity, indicating that there's a negligibly small chance of having an infinitely small value of x, and it is 1 when x goes to plus infinity, which says that there is a 100% chance of observing some value. The midpoint, where $C(x) = 0.5$ is the median.

Practical Considerations

Once you've collected data and plotted a histogram, how do you transform this into a pdf. If histogram bin i contains n_i points, then the fraction n_i/N will tell you the probability that an observation appears in that bin. We also need to divide by the bin width Δx, so that the pdf will integrate to one. Thus the empirical pdf has bins of height $n_i/(N\Delta x)$. Figure 1 shows sample pdfs for wind data.

Observational data often have *Gaussian* or *normal* distributions—that's the classic bell-shaped curve that professors sometimes use to fix grades—and most statistical theory assumes that quantities are normally distributed, with

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]. \tag{4}$$

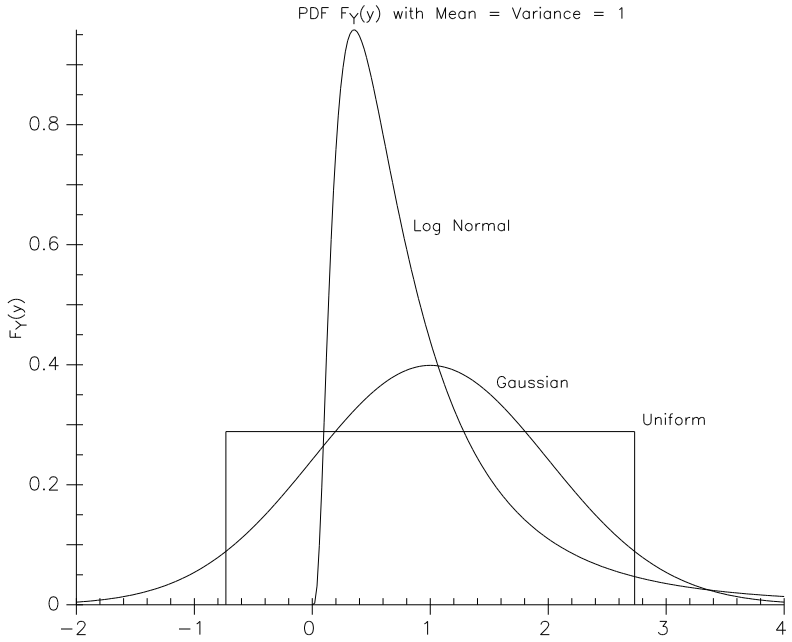

Figure 2: Examples of probability density functions with unit mean and variance.

where σ is the standard deviation. The corresponding cumulative distribution function is the error function. These analytic forms are used to derive much of the basic statistical theory that underlies data analysis.

However, other forms of pdfs often appear in observations. Figure 2 shows some common sample pdfs. Velocities, both in the ocean and in the atmosphere, sometimes appear more double exponential than Gaussian.

$$P(x) = \frac{1}{\sigma\sqrt{2}} \exp\left[-\frac{|x|\sqrt{2}}{\sigma}\right]. \tag{5}$$

Speeds, since they are always positive, roughly follow a Rayleigh distribution (not shown but much like the log normal distribution). Wind directions can be nearly a uniform distribution in cases where the wind is equally likely to blow in any direction between 0 and 2π.

Do two distributions differ?

PDFs are often used only as a concept to help teach statistics and to help explain the concepts of mean, variance, skewness, and kurtosis. But it's natural to ask how you can use them not only as a tool to help you learn about statistics, but also as a tool to help you learn about the ocean.

One key question is to ask whether PDFs in two regions are the same or different. In other words, what is the probability that two random sets of data were drawn from the same underlying distribution. *Numerical Recipes* provides a cogent description of this topic.

One strategy is to use a Kolmogorov-Smirnov test, in which you compute the maximum separation D between the empirical cumulative distribution function and either a known (theoretical or analytic) cumulative distribution function or else a second empirical cumulative distribution function. In essence, if the difference is small, the PDFs are potentially drawn from the same data set. If the difference is large, the data are inconsistent with the null hypothesis that the data come from the same distribution. The trick is to decide what represents a large difference, and that's what the Kolmogorov-Smirnov statistic aims to provide. *Numerical Recipes* explains how to compute this, and Matlab has a usable function, so I won't go into it here. However, a few comments are in order. The Kolmogorov-Smirnov test has a reputation for always failing. That's partly because we often have a poor estimate of the number of degrees of freedom N_{eff} in our data.

If N_{eff} is less than N, the number of data points, then we'll set the wrong standard for the test. You might imagine that you could fix the K-S test by simply plugging N_{eff} into the equations, but that doesn't work, because the correlated data are smoothly varying. You're better off decimating the data to yield a number of samples consistent N consistent with N_{eff}.

A second strategy is to use a χ^2 test to evaluate your empirical PDFs. For comparisons of observed with theoretical PDFs, the χ^2 statistic is

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}, \tag{6}$$

where N_i is the observed number of events in bin i, and n_i is the theoretical or expected number of events in bin i. For comparisons between two distributions,

$$\chi^2 = \sum_i \frac{(N_i - M_i)^2}{N_i + M_i}, \tag{7}$$

where N_i and M_i are each observed numbers of events for bin i. The values of χ^2 are evaluated using the χ^2 probability function $Q(\chi^2|\nu)$, where ν is the number of bins (or the number of bins minus one, depending on normalization).

Extreme events

If you're looking at PDFs, you might also want to think about the stochastic differential equation that describes the evolution of the PDF. Consider a system that follows the equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}(\mathbf{x}) + \mathbf{B}(\mathbf{x})\eta, \tag{8}$$

where $\mathbf{x}$ is a state vector, $\mathbf{A}(\mathbf{x})$ describes the drift of the system, η is noise, and $\mathbf{B}(\mathbf{x})$ describes how the noise depends on the state vector $\mathbf{x}$. The Fokker-Planck equation that describes the evolution of the PDF of this system is

$$\frac{\partial p(\mathbf{x},\mathbf{t})}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} A_i p(\mathbf{x},t) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} (\mathbf{B}\mathbf{B}^T)_{ij} p(\mathbf{x},t). \tag{9}$$

Without worrying about the details of the Fokker-Planck equation, you can use still ask about the drift $\mathbf{A}(\mathbf{x})$ and the matrix $\mathbf{B}$. In basic terms, you can estimate them:

$$\mathbf{A}(\mathbf{x}) = = \lim_{\Delta t \to 0} \frac{1}{\Delta t}(\mathbf{X}(\mathbf{t}+\delta\mathbf{t}) - \mathbf{X}(\mathbf{t})) \tag{10}$$

$$\mathbf{B}(\mathbf{x})\mathbf{B}^{\mathbf{T}}(\mathbf{x}) = = \lim_{\Delta t \to 0} \frac{1}{\Delta t}(|\mathbf{X}(\mathbf{t}+\delta\mathbf{t}) - \mathbf{X}(\mathbf{t})|)(|\mathbf{X}(\mathbf{t}+\delta\mathbf{t}) - \mathbf{X}(\mathbf{t})|)^{\mathbf{T}}. \tag{11}$$

For a single time series, this means that you'll bin average as a function of x. If $B(x)$ turns out to be constant as a function of x, this will tell you that noise is additive, which is what we typically assume. But if $B(x)$ is not constant, that may be a sign of multiplicative noise—in other words, the amplitude of η depends on the state of the system. That has important implications for how we think about physical systems.

Another interesting test is to plot the kurtosis of your data as a function skewness. This too can yield some perspective on the multiplicative character of noise implicit in observations. For more on this topic, you can consult the textbook by Gardiner (*Handbook of Stochastic Methods for Physics, Chemistry and the Natural Science*). Philip Sura (now at Florida State University) has written a number of papers that explore stochastic methods as applied to oceanographic data.