

Figure 1: Observed probability distribution functions of wave height H , wave period T , and wave steepness s for breaking waves (br), nonbreaking waves (nb) and all waves (all), from Holthuisen and Herbers, *J. Phys. Oceanogr.*, **16**, 290-297, 1986.

More on Probability Density Functions: Applications, Details, and Examples

PDF Examples

We have seen that probability density functions are nice as a more universal form of a histogram, but what are they really used for? Here are a couple of concrete examples.

Figure 1 shows probability density functions for waves measured in the North Sea. The investigators were interested in understanding differences between breaking waves and nonbreaking waves. These differences are important for a variety of reasons. Breaking waves are white caps, so they have lots of foam and may matter for understanding how the ocean takes up gas. Breaking waves are also fairly energetic and may matter for predicting how well oil platforms will survive or for deciding the likelihood that a ship will be lost at sea. Although breaking waves are important for our understanding of the ocean and physically interesting in their own right, they are hard to study directly. These results show that the PDFs for breaking and nonbreaking waves differ. However, in this study, the authors concluded that the PDFs overlapped too much to allow a simple distinction between breaking and nonbreaking waves on the basis of wave height, wave period, or steepness alone.

Figure 2 shows an example of seismic risk assessments carried out by civil engineers or seismologists to evaluate the risk of a major earthquake occurring during a specified time interval. Here planners try to reduce the risk so that within the next 50 years, there is less than a 2% chance of experiencing significant ground motion due to an earthquake. That's roughly equivalent to saying that the likely interval before the next earthquake will be 2500 years (since 2% of 2500 years is 50 years). Note that the PDFs in panels a-c

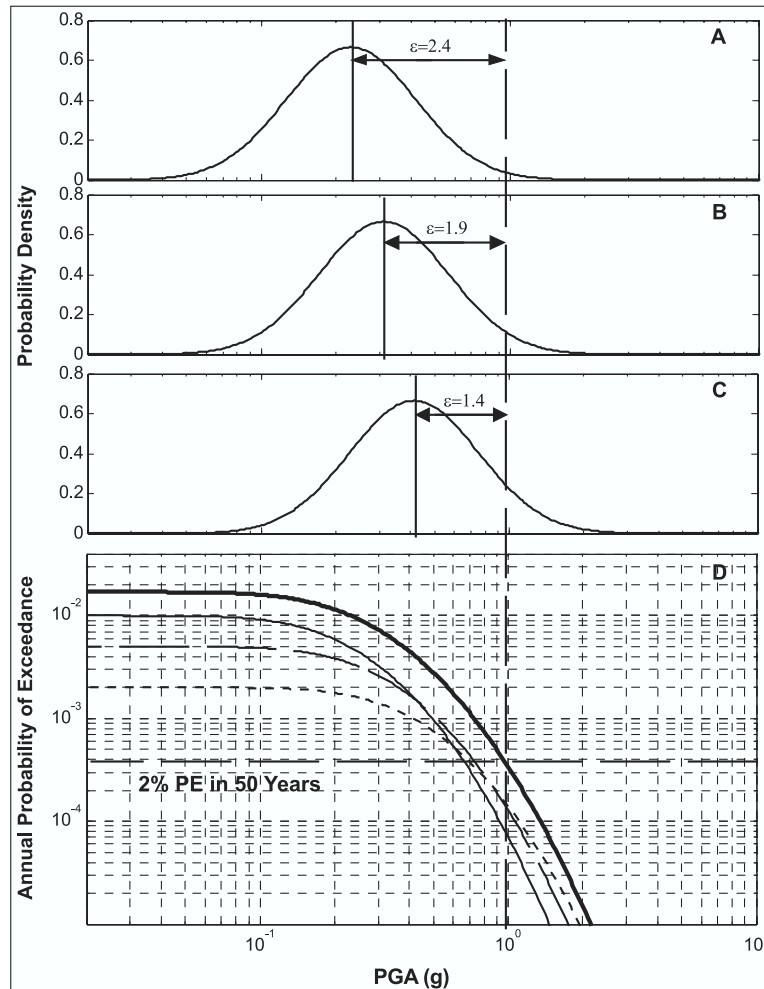


Figure 2: Probability density functions used to predict seismic hazards. The x-axis, PGA, indicates the probability of a gravitational acceleration, and values exceeding $0.97g$ are considered a risk. Panels a-c show log normal distributions for earthquakes at three possible locations, each assumed to be a different strength. Panel (d) shows the probability of having an event with an acceleration exceeding the value on the x-axis for each of the three separate locations (light lines) and for the sum of all three (thick black line). From Wang and Ormsbee, *Eos Trans. Amer. Geophys. Union*, **86**, pp. 45, 51-52, 2005.

look like Gaussian or normal distributions, but the x-axis is logarithmic. With a linear axis, we would see that the PDFs are skewed with long positive tails containing infrequent extreme events. Similar analyses are used to predict the probability that a river will experience a 100-year flood, for example.

Broader Applications

Since we can define functional forms for PDFs, this makes them useful as an analytic tool for predicting statistical results. One important property of the pdf is that we can use it to determine the mean and higher moments.

$$\langle x \rangle = \int_{-\infty}^{\infty} xP(x) dx. \quad (1)$$

Why is this? Suppose I want to compute the mean of x from a set of observations. I can do it by adding up all the measurements of x and dividing by N . Alternatively, I could sort the observations by values of x .

Each possible data value x_j would have a corresponding number of observations n_j . Thus my estimate

$$\bar{x} = \frac{1}{N} \sum_{j=1}^J x_j n_j = \frac{1}{N \Delta x} \sum_{j=1}^J x_j n_j \Delta x. \quad (2)$$

If I take the limit as Δx becomes arbitrarily small, this becomes the expression in (1):

$$\langle x \rangle = \lim_{J \rightarrow \infty} \frac{1}{N} \sum_{j=1}^J x_j n_j = \int_{-\infty}^{\infty} x P(x) dx. \quad (3)$$

The same formulation applies if we want to compute the second moment, $\mu_2 = \langle (x - \langle x \rangle)^2 \rangle$ or any higher moment.

$$\mu_n = \langle (x - \langle x \rangle)^n \rangle = \int_{-\infty}^{\infty} (x - \langle x \rangle)^n P(x) dx. \quad (4)$$

Using this formulation, for a Gaussian $\mu_4 = 3\mu_2^2$, so that the kurtosis $\mu_4/\mu_2^2 = 3$. Similarly, we can show that the skewness $\mu_3/\mu_2^{3/2}$ is zero for a Gaussian—not a surprising result since a Gaussian distribution is completely symmetric.

Here's an example of how to compute moments of a theoretical PDF. Consider the uniform distribution:

$$P(x)dx = \begin{cases} dx & \text{for } 0 < x < 1 \\ 0. & \text{otherwise} \end{cases} \quad (5)$$

First we check that the PDF is defined correctly:

$$\int_{-\infty}^{\infty} P(x) dx = \int_0^1 dx = x|_0^1 = 1 \quad (6)$$

which is exactly what we want. The mean of data from this PDF is:

$$\langle x \rangle = \int_{-\infty}^{\infty} x P(x) dx = \int_0^1 x dx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2}. \quad (7)$$

The variance is

$$\langle (x - \langle x \rangle)^2 \rangle = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 P(x) dx = \int_0^1 (x - \langle x \rangle)^2 dx = \int_{-1/2}^{1/2} y^2 dy = \frac{y^3}{3} \Big|_{-1/2}^{1/2} = \frac{1}{12}. \quad (8)$$

And we can continue on. The third moment is:

$$\mu_3 = \langle (x - \langle x \rangle)^3 \rangle = \int_{-\infty}^{\infty} (x - \langle x \rangle)^3 P(x) dx = \int_0^1 (x - \langle x \rangle)^3 dx = \int_{-1/2}^{1/2} y^3 dy = \frac{y^4}{4} \Big|_{-1/2}^{1/2} = 0, \quad (9)$$

so the skewness is 0. The fourth moment is:

$$\mu_4 = \langle (x - \langle x \rangle)^4 \rangle = \int_{-\infty}^{\infty} (x - \langle x \rangle)^4 P(x) dx = \int_0^1 (x - \langle x \rangle)^4 dx = \int_{-1/2}^{1/2} y^4 dy = \frac{y^5}{5} \Big|_{-1/2}^{1/2} = \frac{1}{80}, \quad (10)$$

so the kurtosis is $\mu_4/\mu_2^2 = 144/80 = 1.8$.

After all of this discussion, you might think that it would make sense to compute the mean of your observations by first deriving the empirical pdf and then numerically integrating the pdf multiplied by x . This would be a mistake, since these procedures would introduce numerous errors into your results.

Next time we'll focus on the central limit theorem which provides a formal framework for explaining why data so often have Gaussian distributions.