A short note on the Y-axis on CDFs, PDFs and PMFs. (Humphrey 2021)

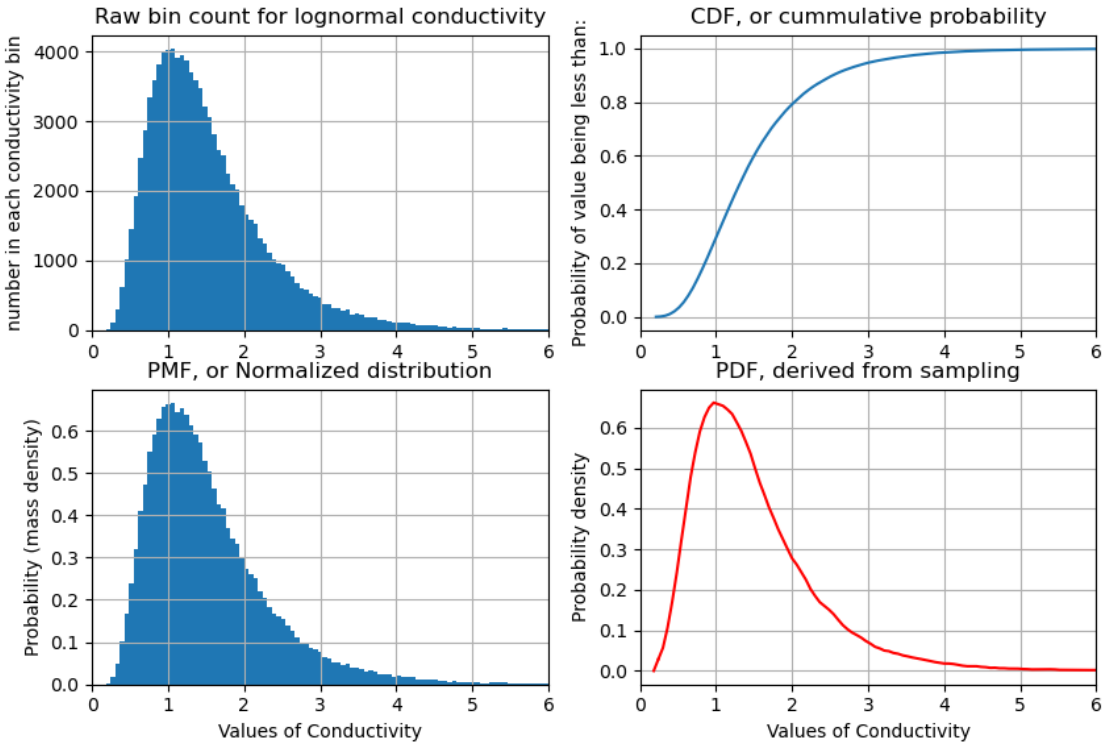
Probability is defined as a number between 0 and 1, which describes the likelihood (note the circular definition!) of the occurrence of some particular event in some range of events. 0 means an infinitely unlikely event, and 1 means the certain event. The term 'event' is very general, but for us can usually be thought of as one sample data point in an experiment, and the 'range of events' would be all the data we would get if we sampled virtually an infinite data set on the experiment. The variability of the samples could be due to error, real-world heterogeneity, process variability etc. however this is not important. We are only interested in quantifying the sample variability.

A set of measured (or sampled) values from some real source of data, or a set of samples from a theoretical source of numbers can be regarded as samples that can reveal the actual character of the source. For example, if the samples are real data measurements, then their distribution can estimate the actual errors associated with the data source. A real-world example would be: if we take a 'lot' of samples, or conductivity estimates, from a rock formation that has a 'log-normal' distribution of conductivities, then the sample histogram reveals the shape of the distribution that these samples came from (illustrated by the upper left of the plot group below). Since we only have samples, we can use the following method to get an estimate of the full data distribution. The samples are placed into evenly spaced bins and plotted as a histogram. If we assume the samples reveal the underlying real data behavior, then we can interpret the histogram as a distribution of the probabilities of sampling the underlying real data. Except, the y-axis in the histogram is not a probability axis, it is just a raw number of samples per bin.

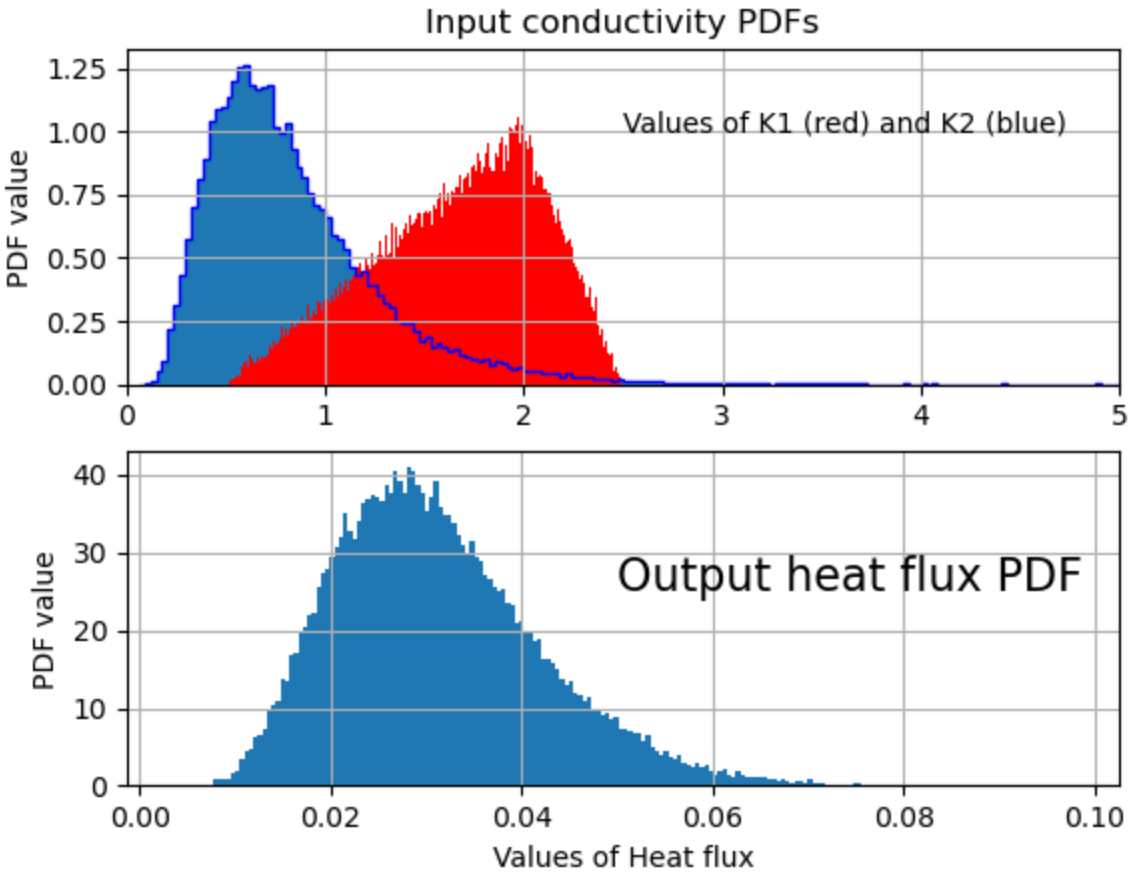
The histogram can be normalized so that the y-axis value becomes an approximate estimate of probability. Normalization makes the y-axis take values that will make the total area under the histogram equal to 1, which is saying that the total probability of the entire distribution of events is equal to the certain event or to a probability of 1. To normalize a histogram to reduce to a probability, the y-values of the bins need to be divided by the total number of samples and by the width of each bin. The y-axis in the PMF plot, probability mass density, give the probability of sampling a value (x-value) that would fit into a particular bin, but we need to multiply by the bin width. (PMFs only have value if this width information is included.)

Once we have an estimate of the distribution as illustrated by the bottom left plot, it is straightforward to integrate this curve and make an estimate of the Cumulative probability Distribution function (CDF) as illustrated in the top right-hand plot. The y-axis in this plot (CDF) is interpreted as the probability of any event less than the x-axis value. For example, the probability of sampling any conductivity less than 2 is about 0.8 or 80%. Notice that the probability of any event less than the maximum value is 1, which implies the 'certain event'. The final step in producing a probability estimate of the underlying real data is to take the derivative of the CDF with respect to the x-axis, which is the Probability Density Function (PDF), illustrated in the lower right-hand plot. The PDF is the most common way of illustrating the

characteristics of a distribution. It is important, and not obvious, to understand the y-axis in the PDF. The y-axis (probability density) does not give the probability of an x-axis value! The probability density gives the probability of the events in a **span** of the x-axis. To illustrate, the probability of an event being between 2 and 2.5 (lower right-hand plot) is given by the area under the PDF plot between the x-axis values of 2 to 2.5, which yields a probability of about 0.09 or 9%. A corollary is that the probability of an exact value on the x-axis is always 0, even though the PDF density value may be large, this is because the area under the curve is zero.



Cautionary note: it may not be intuitive that for PDFs, the PDF probability densities can easily be more than 1 if the x-axis range is small. See the lower plot below, where the 'Probability density' can be large, such as 40 for an x-axis value of 0.03. The calculated probabilities from the PDF however will always remain less than 1, since the integral is basically the y-axis value times the x-axis span (which in this case is very small).


Input conductivity PDFs
PDF value
1.25
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
Values of K1 (red) and K2 (blue)
PDF value
40
30
20
10
0
0.00
0.02
0.04
0.06
0.08
0.10
Values of Heat flux
Output heat flux PDF