Distributions of continuous variables - Notes for Year 12

Dr Richard Kenderdine

Continuous variables are those that are measured eg height, weight, time, rainfall. To produce any graphical sumaries of a continuous variable the data must be first grouped into classes.

This note looks at the theoretical probability models used to analyse continuous data.

We use X to represent the variable and x to represent a realisation (numerical observation) of that variable. For example, the probability that the amount of rainfall (X) was less than a specified value of x = 50mm would be expressed as $P(X < 50)$.

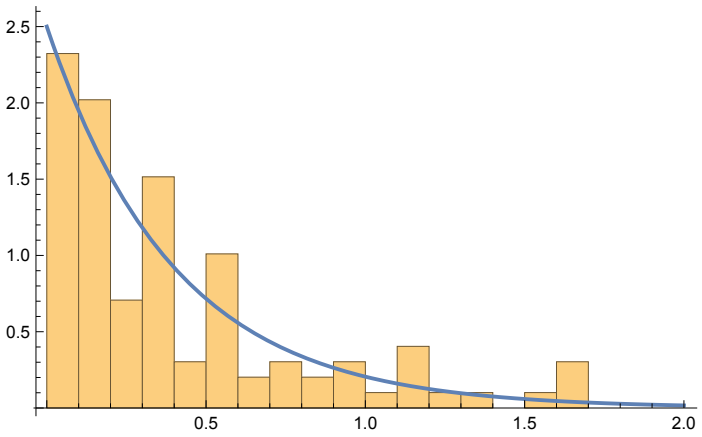
Probability density function (pdf)

The probability density function (pdf) is a theoretical function that corresponds to the relative frequency polygon from sample data (if we have population data then the pdf is the relative frequency polygon). The pdf can take any form with the constraints that it is a non-negative function (no part lies below the x-axis in the domain) that integrates to 1 over some interval [a, b].
For example, all the following functions integrate to 1 with the given limits:

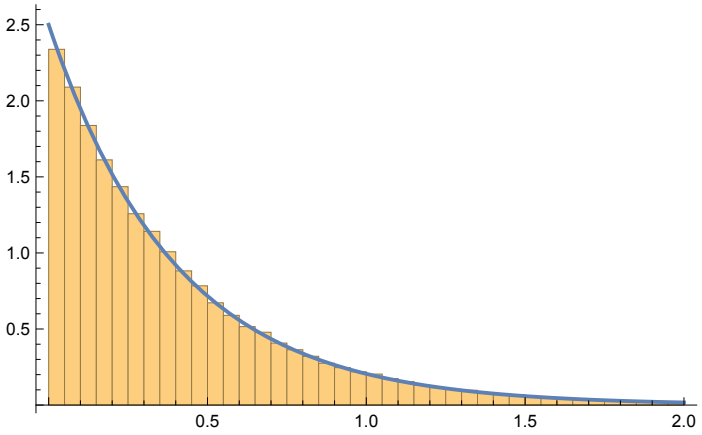
$$\int_1^{10} f(x)\,dx = 1 \qquad \int_0^{\infty} f(x)\,dx = 1 \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

Examples in this note use the exponential distribution that is used to model the life of electronic components and other items. The pdf is $k\,e^{-k\,x}$ for some parameter value k and $\int_0^{\infty} k\,e^{-k\,x}\,dx = 1$.

This plot shows the pdf together with a histogram summarising a sample of 100 observations drawn from a population that follows an exponential distribution with k = 2.5



The histogram only approximates the pdf. However when the sample size is increased to 100000 we see a much better result:

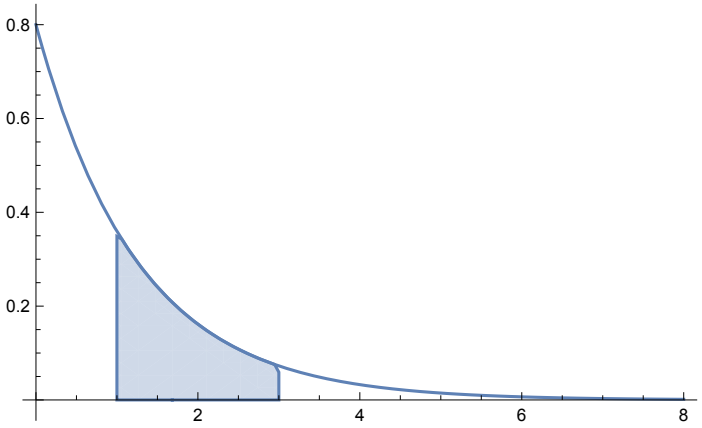


The pdf allows us to determine the theoretical probability of obtaining observations betwwen two values (with continuous data the probability of obtaining a single value is 0).
We use a definite integral:

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)\, dx$$

to find the probability that the value of the random variable will lie between x_1 and x_2.

Here is the pdf of an exponential distribution with parameter 0.8. The area bounded by the curve, the x-axis and the limits x = 1 and 3 is shaded:



The area is found from $\int_1^3 0.8\, e^{-0.8x}\, dx = 0.3586$

Hence $P(1 \le X \le 3) = 0.3586$

Cumulative distribution function (cdf)

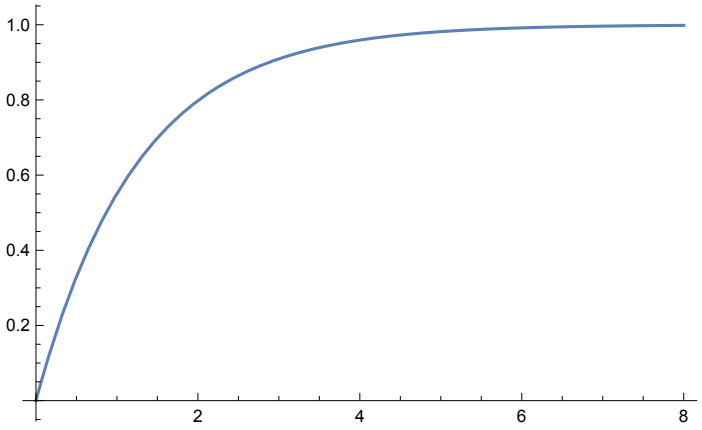
The cumulative distribution function is the primitive function of the pdf. It corresponds to the cumulative relative frequency polygon for sample data (or is the cumulative relative frequency polygon for population data).

To evaluate the cdf at x we use

$$F(x) = \int_a^x f(t)\, dt \qquad \text{for } a \le x \le b$$

Upper case F refers to the primitive funtion for $f(x)$. The cdf has the properties $F(a) = 0$ and $F(b) = 1$. Thus $F(x)$ is the area under the pdf from the lower limit of the domain to x.

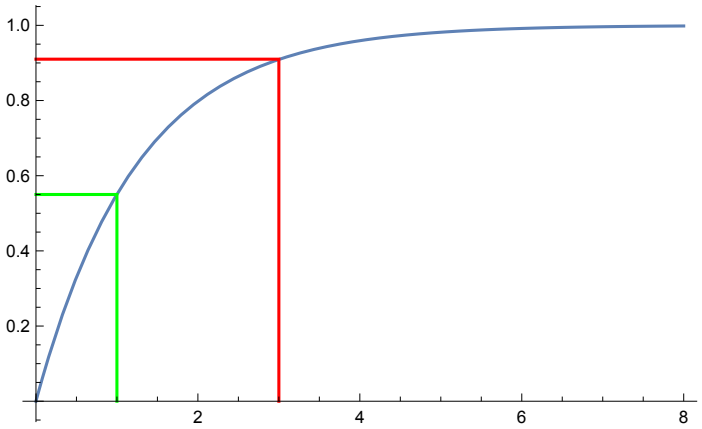
Here is the cdf for the exponential distribution with parameter 0.8:



For the previous example of the exponential distribution with parameter 0.8, the cdf is $1 - e^{-0.8x}$ and so we find $F(1) = 0.5507$ and $F(3) = 0.9093$. This means $P(X < 1) = 0.5507$ and $P(X < 3) = 0.9093$.

Hence $P(1 \le x \le 3) = 0.909 - 0.551 = 0.3586$.

Here is the cdf showing $x = 1$ and 3 together with the function values 0.5507 (green) and 0.9093 (red) respectively:



So if we have the cdf, either as a graph, a table of values or as a function, we can determine the probability without integrating the pdf..

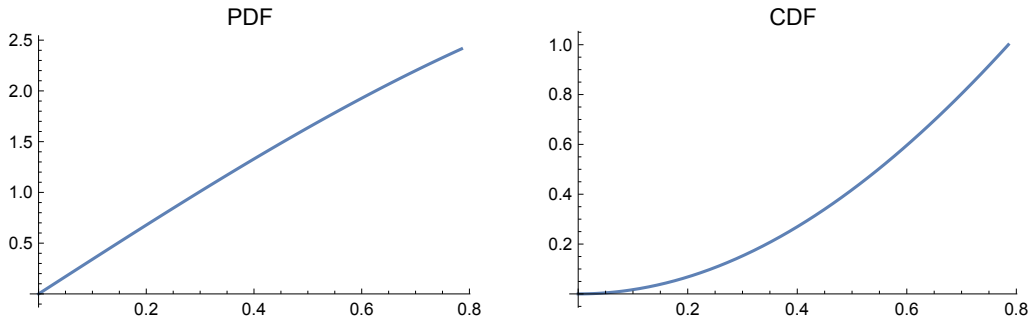
Finding the quartiles

Consider the PDF $c\sin(x)$ for $0 \leqslant x \leqslant \frac{\pi}{4}$ with a constant c to be determined so that $\int_0^{\frac{\pi}{4}} c\sin(x)\,dx = 1$

We find $\int_0^{\frac{\pi}{4}} c\sin(x)\,dx = c\left(1 - \frac{1}{\sqrt{2}}\right) = c\left(\frac{2-\sqrt{2}}{2}\right) \Longrightarrow c = \frac{2}{2-\sqrt{2}} = 2 + \sqrt{2}$ to give the value of the integral to be 1..

Hence the PDF is $\left(2 + \sqrt{2}\right)\sin(x)$

The CDF is then $\left(2 + \sqrt{2}\right)\int_0^x \sin(t)\,dt = \left(2 + \sqrt{2}\right)[1 - \cos(x)]$



To find the quantiles we need to find x such that $\left(2+\sqrt{2}\right)[1 - \cos(x)] = q$ for q = 0.25, 0.5 and 0.75

Re-arrange to give $x = \cos^{-1}\left[1 - \frac{q}{2+\sqrt{2}}\right]$ (using radians).

We have the results

Quartile	x
Q1 (q = 0.25)	0.385
Q2 (q = 0.5)	0.548
Q3 (q = 0.75)	0.676

These values are shown in the plot of the CDF: