Vector Calculus and Multiple Integrals

New material on statistical distributions

Rob Fender, 2019

In 2019 a new, relatively small, component was added to the syllabus for CP4, namely an introduction to statistical distributions. In HT2019 the material was covered in a single lecture and via the introduction of a couple of questions to the problem sheets.

These notes are a summary of the new material as presented in HT2019 and can be considered as an addition to the older course notes.

An introduction to statistical distributions

Continuous distributions, PDF and CDF

The statistical distributions we consider in this small addition to CP4 in 2019 are of the form where a variable X may be drawn from a population described by a statistical distribution. This *continuous random variable* X (discrete distributions also exist) has a range in the form of an interval or a union of non-overlapping intervals on the real line (possibly the whole real line). The variable x is an instance of this distribution. Note that we can only evaluate the probability of x taking some value in a given range; formally the probability of X taking some precise value x is zero: P(X=x) = 0.

The probability of x taking some value in the range x to x+dx is given by the integral of the *Probability Density Function* (PDF) over that range. The *Cumulative Distribution Function* (CDF) as a function of x is the integral of the PDF from zero to the value of x. The following formal statement relates the CDF and PDF:

Consider a continuous random variable X with an absolutely continuous CDF $F_X(x)$ *. The function* $f_X(x)$ *defined by* $f_X(x)=\frac{d(F_X(x))}{dx}$ *(if* $F_X(x)$ *is differentiable at x) is called the probability density function (PDF) of X.*

In other words, the PDF is the derivative of the CDF with respect to x and the CDF is the integral of the PDF between zero and x.

Example: your friend tells you that she will stop by your house sometime after or equal to 1 p.m. and before 2 p.m., but she cannot give you any more information as her schedule is quite hectic. Your friend is very dependable, so you are sure that she will stop by your house, but other than that we have no information about the arrival time. Thus, we assume that the arrival time is completely random in the 1 p.m. and 2 p.m. interval.

1. What is the sample space S?

The sample space is simply S = [1,2)

(note the use of "[" to indicate equal to or greater than and ")" to indicate less than)

2. Sketch the PDF and CDF

The PDF is a uniform top hat function between 1 and 2pm, and zero outside. The CDF is the integral of this PDF to x, and as such is a slope from 0 at 1pm to unity at 2pm. The height of the PDF is determined such that this total integrated area is 1 (since a probability of 1 means that the event *must* happen in that interval).

3. What is the probability of P(1.5)? Why?

P(1.5) = 0, as does P(X=x) for any specific value, since for a continuous probability function, the probability has to be integrated over some interval.

4. What is the probability of T ∈ [1,1.5)?

This is equal to the value of the CDF at X = 1.5, which we can see intuitively will be 0.5. In other words there is a 50% chance that your friend will arrive in the first half of the agreed interval (since the distribution is uniform between 1 – 2pm).

5. For $1 \le a \le b \le 2$, what is $P(a \le T \le b) = P([a,b])$?

P(a,b) = b – a for $1 \le a \le b \le 2$. More generally, for a uniform distribution between limits x1 and x2, the PDF is a top-hat function of height (x2-x1)^{-1} since the total integrated area of the PDF across the sample space = 1. The probability to be in the range (a,b) then corresponds to P(a,b)=(b-a)/(x2-x1), which in this case reduces to b – a since (x2-x1 = 1).

Expectation value and Variance

For a given PDF $f_X(x)$ the expectation value of x (i.e. the mean) is given by

$EX = \int x f_X(x) dx$ and more generally for any function of x is given by $E[g(X)] = \int g(x) f_X(x) dx$

Example: For $f_X(x) = x + \frac{1}{2}$ for 0 ≤ x ≤ 1 and zero otherwise, find $E[X^n]$

Solution:

$$E(X^n) = \int_0^1 x^n (x + \frac{1}{2}) = \frac{(3n+4)}{2(n+2)(n+1)}$$

The Variance of a distribution is given by

$$Var[X] = EX^2 - (EX)^2 = \int x^2 f_X(x) - \mu_X^2 dx$$

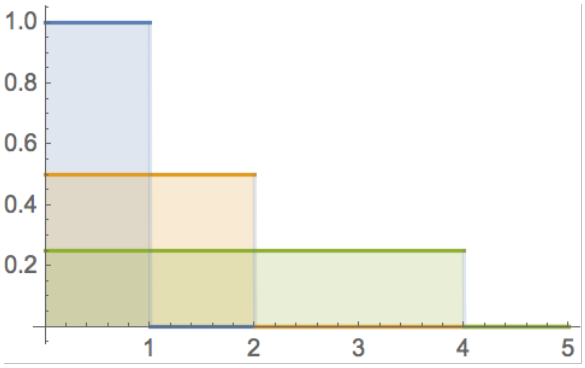
Where μ_X is the mean value (i.e. the expectation value).

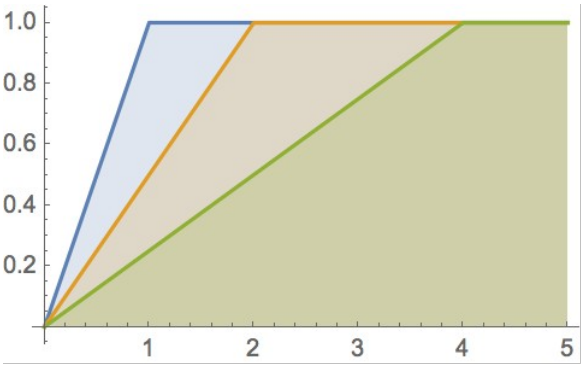
Some common statistical distributions

Uniform distribution (PDF left, CDF right)

$PDF\ f(a<x<b) = constant\ ,zero\ otherwise$

Figures are differing ranges and corresponding constants such that integrated probability = 1

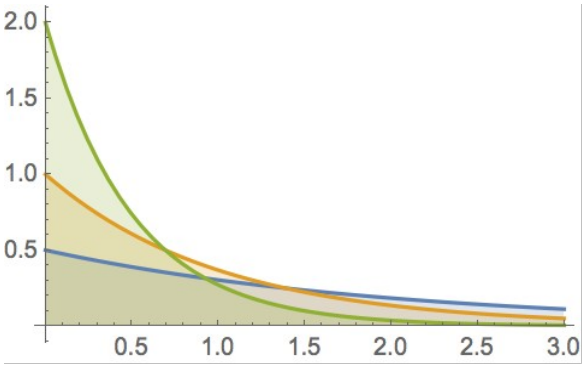


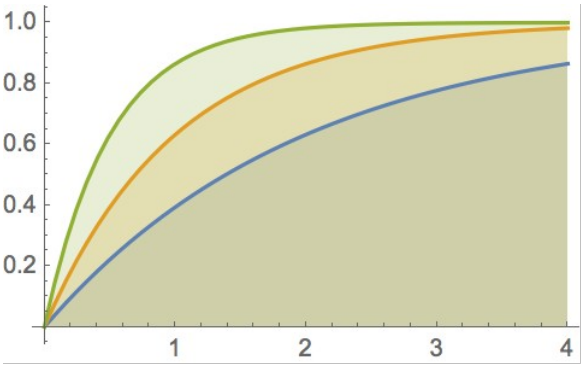


Exponential distribution (PDF left, CDF right)

$PDF\ f(x) = \lambda \exp[-\lambda x]$

Figures are differing values of λ

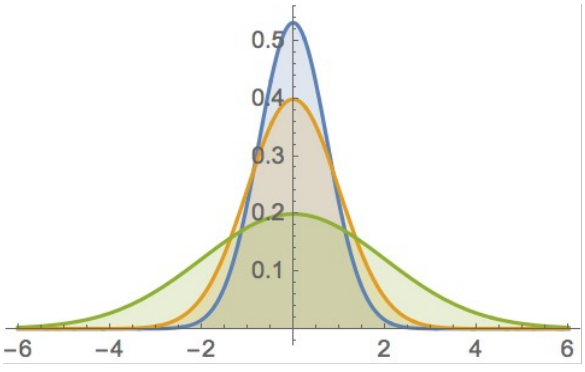


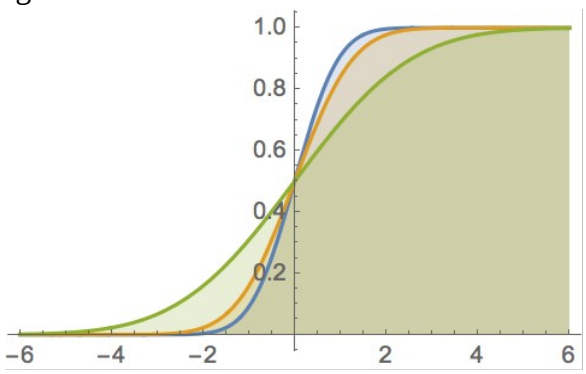


Normal (Gaussian) distribution (PDF left, CDF right)

$PDF\ f(x) = \frac{1}{(\sigma\sqrt{2\pi})} \exp[\frac{-(x-\mu)^2}{2\sigma^2}]$ where μ, σ^2 are the mean and variance respectively.

Figures are Gaussians with zero mean but differing variances.





Joint (2D) probability distributions

We may also consider a 2D joint probability distribution,

$$P((x,y)\in A)=\iint_A f_{XY}(x,y)\,dx\,dy$$

which is normalised in an identical way to the 1D probability distributions

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx\,dy=1$$

Example: For $f_{XY}(x,y)=x+c\,y^2$ for (0 ≤ x ≤ 1 and 0 ≤ y ≤ 1) and zero otherwise, find (a) the constant c and (b) P (0 ≤ X ≤ 1/2 and 0 ≤ Y ≤ 1/2)

Solution

(a) NormaIising, using $\int_0^1\int_0^1 f_{XY}(x,y)\,dx\,dy=1$ (which is the same as above but now we know the sampled space is restricted to the range 0 – 1 for each variable), we get c = 3/2.

(b) This is a straightforward 2D integral, making use of the value of c we found in (a)

$$\int_0^{1/2}\int_0^{1/2}\left(x+\frac{3}{2}y^2\right)dx\,dy=\frac{3}{32}$$

We may also *marginalize* our multi-dimensional PDF in order to evaluate it simply in terms of a single variable:

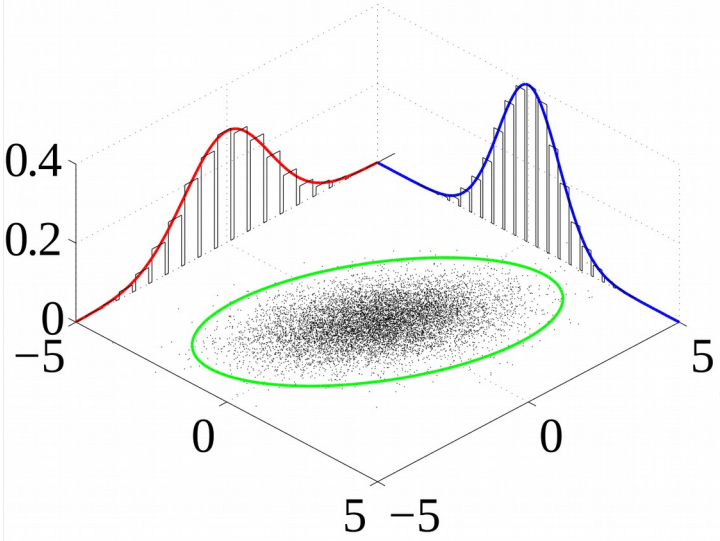
$f_X(x)=\int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy$ for all x, and $f_Y(y)=\int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx$ for all y.

Example: Find the marginal PDFs $f_X(x)$ and $f_Y(y)$ for the 2D joint PDF in the problem above.

Solution

$f_X(x)=\int_0^1\left(x+\frac{3}{2}y^2\right)dy=x+\frac{1}{2}$ for 0 ≤ x ≤ 1 and zero otherwise

$f_Y(y)=\int_0^1\left(x+\frac{3}{2}y^2\right)dx=\frac{3}{2}y^2+\frac{1}{2}$ for 0 ≤ y ≤ 1 and zero otherwise



An example of a joint probability distribution measured from data (the points), and some contour of a 2D distribution (likely Gaussian) used to fit them.

The associated marginal probability distributions in x and y are indicated by the blue and red 1D functions respectively.